VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

ALGORITMY PRO VYHLEDÁVÁNÍ TRANSKRIPČNÍCH MOTIVŮ

TRANSCRIPTION MOTIF FINDING ALGORITHMS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

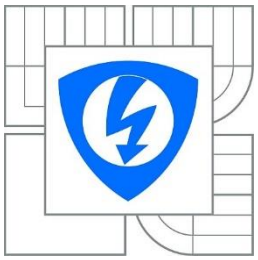
AUTOR PRÁCE — BARBORA PEŘINOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Barbora Peřinová ***ID:*** 154646
Ročník: 3 ***Akademický rok:*** 2014/2015

NÁZEV TÉMATU:

Algoritmy pro vyhledávání transkripčních motivů

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s procesem transkripce DNA sekvencí a významem transkripčních motivů. 2) Vypracujte rešerši existujících algoritmů pro automatické vyhledávání motivů. Zaměřte se především na pravděpodobnostní metody a metody založené na vyhledávání slov. 3) Pro jeden z algoritmů navrhněte vývojový diagram a pseudokód. 4) Vybraný algoritmus implementujte v libovolném programovém prostředí. 5) Realizovaný algoritmus validujte na souboru uměle vytvořených sekvencí s vnesenými motivy a následně na genu kvasinky S. cerevisiae. 6) Diskutujte získané výsledky.

DOPORUČENÁ LITERATURA:

[1] DAS, M. K. a DAI, H. K. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007, 8, S21.
[2] HU, J., LI, B. a KIHARA, D. Limitations and potentials of current motif discovery algorithms. Nucleic Acids Research. 2005, 33, 4899-4913.

Termín zadání: 9.2.2015 ***Termín odevzdání:*** 29.5.2015

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt:

Tato bakalářská práce se zabývá vyhledáváním transkripčních motivů v sekvencích DNA. V první části je popsán proces transkripce DNA, význam transkripčních motivů a je uveden přehled databází transkripčních motivů. Druhá část obsahuje rozdělení algoritmů pro vyhledávání transkripčních motivů a detailně popisuje sedm existujících algoritmů. V praktické části byly vytvořeny funkce pro analýzu podle algoritmu Oligo-Analysis. Vytvořené funkce byly validovány na uměle vytvořených sekvencích s vnesenými motivy. Dále byla provedena analýza dvou rodin koregulovaných genů a výsledky porovnány s hodnotami, kterých dosáhli autoři algoritmu Oligo-Analysis.

Klíčová slova:

DNA, transkripce, transkripční motiv, algoritmy vyhledávání, Oligo-Analysis

Abstract:

This bachelor thesis deals with finding transcription motifs in DNA sequences. The first part describes the DNA transcription process, the significance of transcription motifs and provides an overview of transcription motif databases. The second part contains the classification of transcription motif finding algorithms, and details seven existing algorithms. In the practical part, functions for the analysis according to the Oligo-Analysis algorithm were created. Created functions were validated on the artificially created sequences with introduced motifs. The analysis of two families of coregulated genes was realized and the results were compared with the values achieved by the authors of the Oligo-Analysis algorithm.

Keywords:

DNA, transcription, transcription motif, finding algorithms, Oligo-Analysis

Bibliografická citace

PEŘINOVÁ, B. *Algoritmy pro vyhledávání transkripčních motivů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 50 str. Vedoucí bakalářské práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svou bakalářskou práci na téma „Algoritmy pro vyhledávání transkripčních motivů“ jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne .....

podpis autorky

Poděkování

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce. Dále bych ráda poděkovala celé své rodině za trpělivost a podporu během studia a psaní bakalářské práce.

V Brně dne .....

podpis autorky

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Každý organismus při svém vzniku zdědí od svých rodičů genetickou výbavu, která obsahuje všechny informace pro život a vývoj daného jedince. Expresí genetické informace v organismu vznikají proteiny, které ovlivňují jeho vývoj a současný stav. Velmi důležité je, aby tvorba proteinů měla nějaký řád. V rámci každého organismu tedy dochází k regulaci exprese genetické informace na mnoha úrovních. Na úrovni transkripce se o regulaci starají transkripční faktory, které se váží na specifická místa, transkripční motivy.

Vyhledáváním transkripčních motivů se zabývá mnoho vědců již několik let. Pro většinu organismů je známá pouze část transkripčních motivů a stále se pátrá po dalších. Pro usnadnění pátrání byl vyvinut nespočet algoritmů, které na základě výpočtů či pravděpodobností vyberou ze sekvencí zajímavá místa a ta dále podléhají výzkumu. Znalost transkripčních motivů umožní ovlivňovat expresi genetické informace a takzvaně donutit buňku k tvorbě proteinu v případě, že toho sama není schopna.

Cílem teoretické části této bakalářské práce je seznámit čtenáře s procesem transkripce a její regulací. V další části je popsán význam a rozdělení transkripčních motivů. Porovnání databází transkripčních motivů dostupných na internetu se věnuje třetí část práce. Následující část uvádí rozdělení algoritmů pro vyhledávání transkripčních motivů, detailní popis dnes používaných algoritmů a jejich výhody a nevýhody.

Praktická část práce je zaměřena na realizaci algoritmu Oligo-Analysis v programovém prostředí Matlab. V prvním úseku této části práce jsou popsány všechny vytvořené funkce, jejich vstupní a výstupní hodnoty a způsob jejich použití. Funkce jsou v dalším úseku validovány na uměle vytvořených sekvencích s vnesenými motivy a získané výsledky jsou přehledně uvedeny v tabulkách a okomentovány v textu. Poslední úsek praktické části se zabývá analýzou dvou rodin koregulovaných genů kvasinky *Saccharomyces cerevisiae*. Výsledky této analýzy jsou porovnány s výsledky, kterých dosáhli autoři metody Oligo-Analysis v rámci své studie.

1 TEORETICKÝ ÚVOD

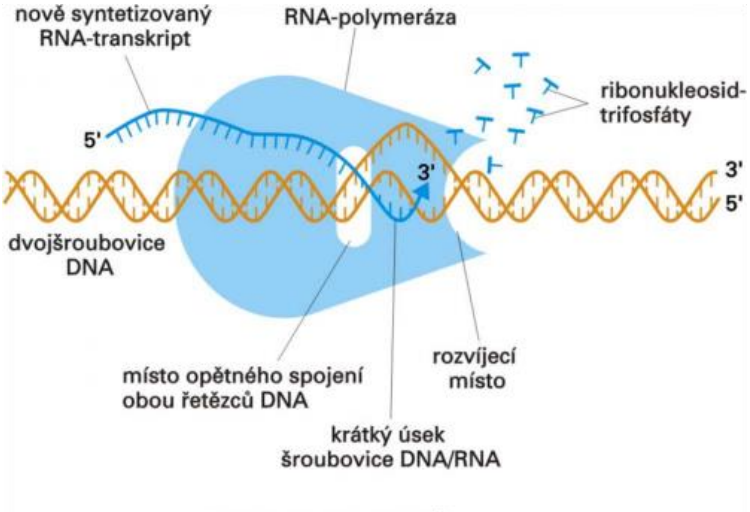
1.1 Transkripce

Genetická informace každého organismu je uložena v sekvenci nukleotidů deoxyribonukleové kyseliny (DNA) v jádře každé buňky. Výjimkou jsou některé viry, které mají genetickou informaci uloženou v sekvenci nukleotidů ribonukleové kyseliny (RNA). Takto jsou v buňkách uloženy informace o primární struktuře proteinů a RNA či o vazbě specifických proteinů k molekule DNA. Pro přečtení této informace musí buňka provést dvě operace, a to transkripci a translaci. V rámci transkripce dochází v jádře buňky k přepisu části DNA na řetězec RNA na základě komplementarity bází. Takto vzniklý primární transkirpt u eukaryotických buněk většinou podléhá posttranskripčním úpravám a je dopraven z jádra buňky do cytoplazmy. Následuje proces translace na ribozomech, který zahrnuje přeložení genetické informace ze sekvence ribonukleotidů do sekvence aminokyselin. Výsledným produktem translace je protein. V případě exprese informace o primární struktuře RNA je samotná transkribovaná molekula RNA hotový produkt, který má určitou funkci v buňce. [1], [2]

Proces transkripce se skládá ze tří částí: iniciace, elongace a terminace. Za počátek iniciace je považováno navázání RNA polymerázy na promotor. RNA polymeráza je enzym, který katalyzuje syntézu RNA pomocí dřeňového enzymu a obsahuje sigma faktor umožňující vazbu enzymu na promotor. Eukaryotické buňky obsahují tři druhy RNA polymerázy, zatímco prokaryotické pouze jeden typ. Promotor je sekvence o délce přibližně 40 nukleotidů, která slouží jako vazebné místo pro RNA polymerázu, signalizuje začátek transkripce a rozhoduje o tom, které vlákno DNA je templátem. Na promotorovou sekvenci se při iniciaci vážou také transkripční faktory, které regulují transkripci a zprostředkovávají navázání příslušné RNA polymerázy na promotor. V případě, že jsou navázány všechny potřebné transkripční faktory a sigma faktor, dochází k částečnému rozpletení řetězců dvoušroubovice DNA a syntéze několika prvních nukleotidů. Iniciace je ukončena uvolněním sigma faktoru. [3], [4]

V průběhu elongace se RNA polymeráza posouvá po templátovém vlákně ve směru od 3‘ konce řetězce k 5‘ konci a syntetizují se příslušné ribonukleotidy (Obrázek 1.1). S pohybem RNA polymerázy dochází k rozvíjení dvojité šroubovice DNA a zpětnému svíjení již transkribované části tak, že je vždy rozvinuto 10 až 20 nukleotidových bází. Již syntetizované ribonukleotidy vytváří DNA/RNA hybrid o délce přibližně 12 nukleotidů, kde je ještě propojena DNA s RNA, na jeho konci se RNA od DNA odděluje. Pokud je v buňce potřeba více určitého proteinu, může být templátové vlákno přepisováno více RNA polymerázami za sebou. [3], [4]

Terminace, neboli ukončení transkripce, nastává přepsáním terminátoru, což je sekvence signalizující konec kódující části. Reakce na terminační signál může probíhat dvěma způsoby. Jednou možnou reakcí na terminační signál je změna struktury RNA transkriptu do podoby vlásenky, což zastaví RNA polymerázu a odštěpí transkript. Druhou možností je navázání Rho faktoru, který rozdělí DNA/RNA hybrid. [3], [4]

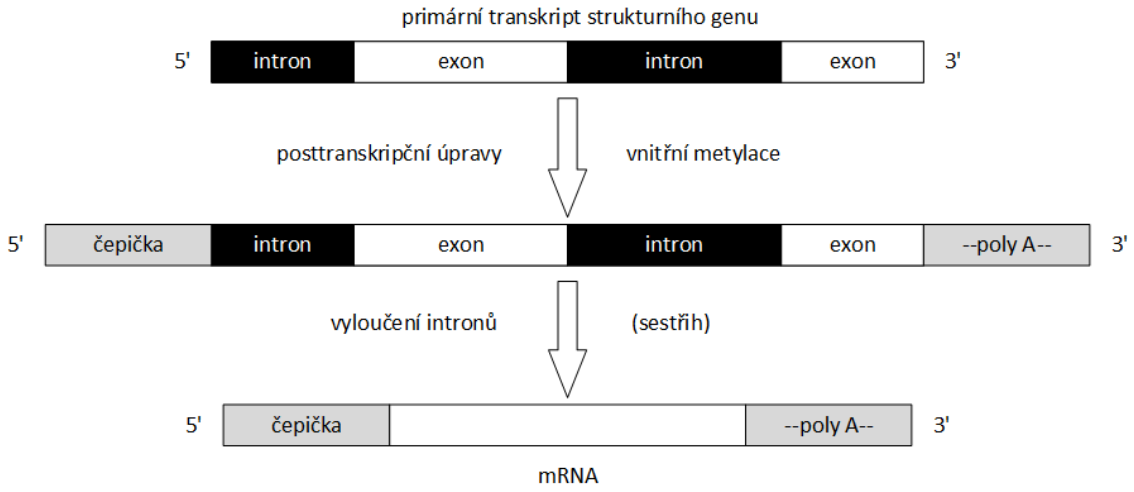


Obrázek 1.1: Průběh elongace [5]

V případě, že je vzniklý primární transkript přepisem strukturního genu eukaryotické buňky, podléhá posttranskripčním úpravám (Obrázek 1.2). Jednou z těchto úprav je připojení čepičky (metylace), kdy se na 5‘ konec transkriptu napojí methylguaninový nukleotid. Na čepičku se ještě naváže specifická bílkovina a tento komplex funguje jako navádějící struktura pro vazbu ribozomu. Podobnou úpravou je polyadenylace, při které dochází k připojení okolo 200 adeninových nukleotidů tvořících ocas na 3‘ konec transkriptu pomocí enzymu polyA-polymerázy. Připojení čepičky a ocasu je ochranou proti rozštěpení transkriptu exonukleázami. Poslední posttranskripční úpravou je sestřih. Primární transkript se skládá z kódujících a nekódujících oblastí. Kódující oblasti nazýváme exony a nekódující introny. Během sestřihu dochází k vystřižení intronů a spojení zbývajících exonů, čímž vzniká funkční mRNA (mediátorová). [6], [7]

1.1.1 Regulace transkripce pomocí transkripčních faktorů

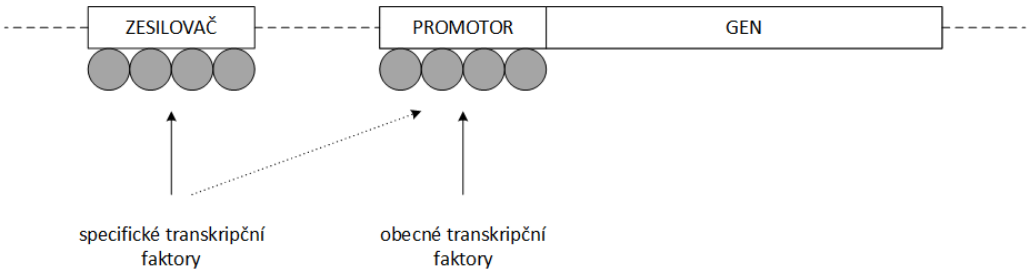
V každém organismu musí docházet k regulaci genové exprese. Pro organismus je zbytečné stále tvořit všechny proteiny, na jejichž tvorbu má geny. Je potřeba, aby každá buňka prováděla expresi takové genetické informace, kterou právě potřebuje. U mnohobuněčných



Obrázek 1.2: Posttranskripční úpravy RNA [6]

organismů s diferenciací buněk do tkání se projevují jen ty geny, které v dané buňce mají nějaký užitek. Například myocyty (svalové buňky) budou exprimovat jiné geny než neurony (nervové buňky). K regulaci dochází na všech úrovních genové exprese, tedy při transkripci, posttranskripčních úpravách i translaci. [6], [8]

Jedním z regulačních prvků jsou transkripční faktory. Transkripční faktory jsou specifické proteiny, které se váží na část sekvence promotoru nebo zesilovače a aktivují či inhibují RNA polymerázu. Sekvence zesilovače je dlouhá v rozmezí 10 a 20 nukleotidů a nachází se v blízkosti genu. Zesilovač nemusí přímo sousedit s promotorem, ale může být vzdálen až tisíce párů bází po směru i proti směru transkripce. Transkripce daného genu je většinou ovlivněna větším počtem transkripčních faktorů, které dohromady tvoří proteinový komplex (Obrázek 1.3). [6], [7]



Obrázek 1.3: Regulace transkripce transkripčními faktory [6]

1.2 Transkripční motivy

Každý transkripční faktor se váže na část sekvence DNA o určitém sledu nukleotidů. Tento konkrétní sled nukleotidů se nazývá transkripční motiv. Transkripční motiv je vždy charakteristický pro jeden určitý transkripční faktor, je poměrně krátký, 5 až 20 nukleotidů, a dochází k jeho opakování v rámci genu nebo i jiných genů. Transkripční motivy se objevují

na obou vláknech DNA. Daný motiv se v určité sekvenci může objevit jednou i vícekrát, ale také se nemusí objevit vůbec. [8]

Podle způsobu umístění motivu v sekvenci nukleotidů rozlišujeme dva druhy motivů, palindromické a vmezeřené. Palindromický motiv je část sekvence, která je shodná se svým reverzním komplementem, například ACATGT. Vmezeřený motiv se skládá ze dvou menších podjednotek, které jsou oddělené mezerou, například ACTXXXGCG, kde X znázorňuje nukleotid mezery. Mezera se vyskytuje vždy ve středu motivu. Transkripční faktory vážící se na vmezeřené motivy jsou složeny vždy ze dvou podjednotek, kde každá z nich má vlastní místo pro kontakt s DNA sekvencí. Tyto podjednotky jsou konzervované, tedy k mutacím u nich dochází v malém množství. Na rozdíl od podjednotek jsou mezery nekonzervované a k mutacím u nich dochází velmi často. Mezery mají většinou pevně stanovenou délku, ale jejich obsah je variabilní. [8], [9]

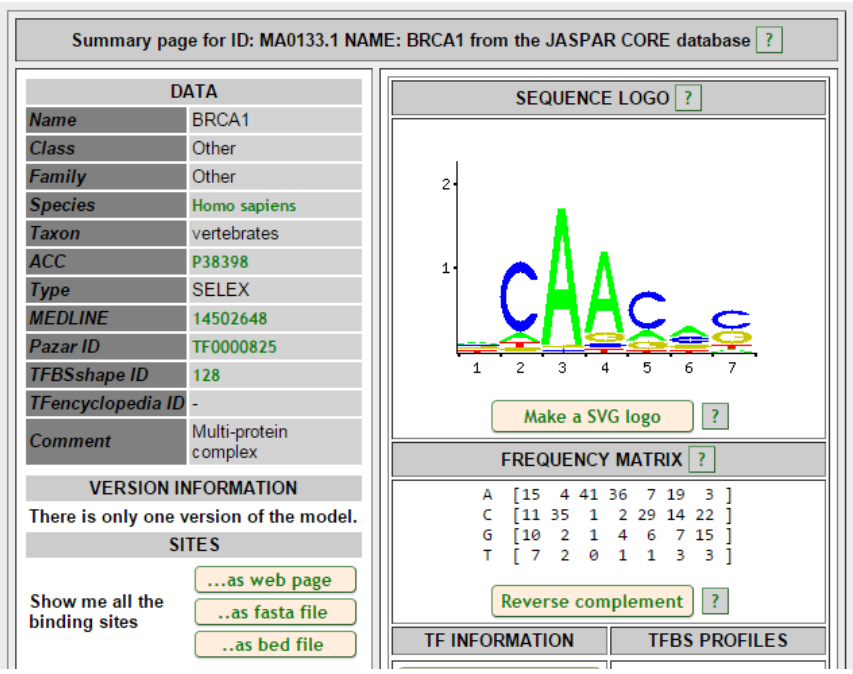
Transkripční motiv je v rámci celé genetické informace velmi malá podjednotka, což ztěžuje jeho nalezení. Sled nukleotidů daného motivu také není vždy stejný, protože dochází k ojedinělým mutacím jednotlivých nukleotidů. Důsledkem těchto faktů je to, že v současné době stále nejsou známy zdaleka všechny transkripční motivy a pořád dochází k hledání nových. Nejlépe jsou prozkoumány jednodušší organismy s méně obsáhlou genetickou informací.

2 DATABÁZE TRANSKRIPČNÍCH MOTIVŮ

V současnosti známé transkripční motivy jsou sepsány v různých databázích přístupných na internetu. Databáze jsou průběžně doplňovány po objevení a ověření nových motivů. Ráda bych se zmínila o třech nejobsáhlejších databázích transkripčních motivů eukaryotických organismů.

2.1 JASPAR

Databáze JASPAR je veřejně dostupná. Je rozdělená na menší samostatné databáze podle druhů organismů. Například JASPAR_CORE Vertebrata obsahuje transkripční motivy pouze obratlovců. Je možné vyhledávat buď v těchto menších databázích nebo přímo v celé databázi JASPAR_CORE. Vyhledávání je možné pomocí ID, názvu či třídy transkripčního faktoru, druhu organismu a typu. ID je unikátní identifikační kód pro každý prvek databáze (formát MAnnnn, kde n je číslo). Vyhledávání je možné pomocí jednoho až tří různých kritérií. Každý záznam obsahuje detailní informace o transkripčním motivu, včetně matice frekvencí výskytu jednotlivých nukleotidů na daných pozicích motivu (Obrázek 2.1). Součástí záznamu je také grafické zobrazení četností nukleotidů, logo motivu. Velkou výhodou je možnost stažení nalezeného motivu ve fasta formátu, který obsahuje všechny variace daného motivu, jaké byly nalezeny. Tyto variace je také možné zobrazit jako webovou stránku. Databáze působí poměrně přehledně a vyhledávání je jednoduché. [10]



Obrázek 2.1: Záznam transkripčního faktoru BRCA1 v databázi JASPAR [11]

2.2 TRANSFAC 7.0

Databáze TRANSFAC je bezplatně dostupná pouze pro studenty, učitele a akademické pracovníky a je potřeba se do ní zaregistrovat. Placená verze je mnohem obsáhlejší než bezplatná a také obsahuje spoustu vedlejších funkcí, například rozšířené nastavení vyhledávání, analýzu nebo možnost stažení dat. TRANSFAC umožňuje vyhledávání šesti různými způsoby. Je možné hledat motivy, jimi ovlivňované geny, transkripční faktory, jiné proteiny interagující s motivy, třídy motivů a matice frekvencí výskytu jednotlivých nukleotidů. Nalezené záznamy jsou pouze v textové formě a není možné jejich stáhnutí v jakékoli formě (Obrázek 2.2). Celá databáze působí velmi nepřehledným dojmem, chybí grafické zobrazení četnosti nukleotidů. Většina popisků je uvedena ve zkratkách a je nutné před použitím prostudovat návod pro alespoň částečnou orientaci uživatele. [12]

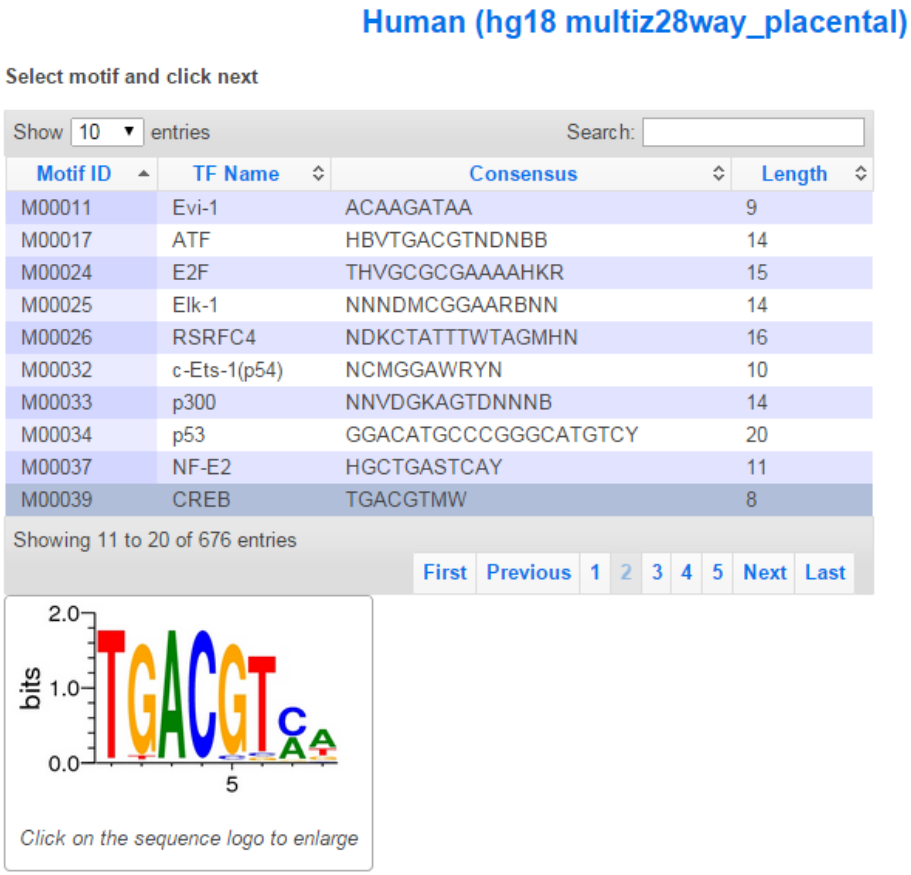
```
AC   T00163
XX
ID   T00163
XX
DT   15.10.1992 (created); ewi.
DT   25.08.2005 (updated); elf.
CO   Copyright (C), Biobase GmbH.
XX
FA   CREB
XX
SY   ATF-47; CREB; CREB-327; CREB-341; CREB-A; CREB-B; CREB1; CREBa
XX
OS   human, Homo sapiens
OC   eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda;
XX
GE   G004624 CREB1; HGNC: CREB1.
XX
HO   ATF, 47-kDa protein, EIIA-EF, EPF, EIIaE-B.
XX
CL   C0008 bZIP; 1.1.2.0.1.1.
XX
SZ   341 AA; 36.7 kDa (cDNA)
XX
SQ   MTMESGAENQQSGDAAVTEAENQQMTVQAQPQIATLAQVSMPAAHATSSAPTVTLVQLPN
SQ   GQTVQVHGVIQAAQPSVIQSPQVQTVQSSCKDLKRLFSGTQISTIAESEDSQESVDSVTD
SQ   SQKRREILSRRPSYRKILNDLSSDAPGVPRIEEEKSEEETSAPAITTVTVPTPIYQTSSG
SQ   QYIAITQGGAIQLANNGTDGVQGLQTLTMTNAAATQPGTTILQYAQTTDGQQILVPSNQV
SQ   VVQAASGDVQTYQIRTAPTSTIAPGVVMASSPALPTQPAEEAARKREVRLMKNREAAREC
SQ   RRKKKEYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD
```

Obrázek 2.2: Záznam transkripčního faktoru CREB v databázi TRANSFAC 7.0 [12]

2.3 MotifMap

Databáze MotifMap obsahuje pouze informace o známých transkripčních motivech modelových organismů (člověka, myši, mouchy, kvasinky a červa). Data přebírá od dvou předchozích databází, tedy JASPAR a TRANSFAC. MotifMap není klasická databáze, jedná se spíše o snahu o zmapování celých genomů jednotlivých organismů. Vyhledávání je možné pomocí tří různých způsobů. Jedním způsobem je vyhledání transkripčního faktoru mezi transkripčními faktory pro daný organismus, zobrazí se logo motivu (Obrázek 2.3).

Další možností je vybrat gen a hledat transkripční motivy před a za místem začátku transkripce. Posledním způsobem je zvolit chromozom, lokaci a délku prozkoumávání. Do tabulky se vypíší motivy vyskytující se v této oblasti, případně v oblastech sousedních. Databáze MotifMap není tak obsáhlá jako dvě předchozí databáze, ale zařadila jsem ji zde kvůli její přehlednosti. Orientace v uživatelském rozhraní je velmi intuitivní, zobrazení výsledků vyhledávání je názorné. [13]



Obrázek 2.3: Vyhledávání transkripčních faktorů pomocí databáze MotifMap [13]

3 ALGORITMY PRO VYHLEDÁVÁNÍ TRANSKRIPČNÍCH MOTIVŮ

Pro vyhledávání transkripčních motivů bylo vyvinuto mnoho různých algoritmů. Většina těchto algoritmů je navržena na vyhledávání motivů v promotorech několika genů, které spadají pod stejnou regulační oblast. V těchto promotorových sekvencích jsou vyhledávány několikanásobně se opakující oblasti, vhodní kandidáti na transkripční motivy. V rámci těchto metod jde vždy o porovnávání sekvencí jednoho genomu. Ukázalo se, že tyto algoritmy fungují poměrně dobře pro kvasinky a jednodušší organismy, ale jejich schopnosti predikce se velmi snížily u fylogeneticky vyšších organismů. Proto došlo k vývoji jiného typu algoritmu, který porovnává genomy více organismů nebo využívá fylogenetického stopování. Základem fylogenetického stopování je předpoklad, že funkční oblasti genomu se vyvíjí značně pomaleji než nefunkční oblasti. Dochází tedy k porovnávání shodně umístěných promotorů v rámci více příbuzných organismů a vyhledávání konzervovaných oblastí (nezměněných vývojem), tedy funkčních oblastí s motivy. Nejnovější algoritmy kombinují oba dva typy vyhledávání, tedy využívají vyhledávání v promotorech v rámci jednoho genomu a také fylogenetické stopování. [8]

Podle způsobu řešení problému můžeme algoritmy pro vyhledávání transkripčních motivů rozdělit na dvě hlavní skupiny: metody založené na vyhledávání slov a pravděpodobnostní metody. Metody založené na vyhledávání slov jsou založeny převážně na složitých výpočtech, využívají například výpočty frekvencí nukleotidů v prohledávané sekvenci a porovnávají je. Tyto metody pracují s celou vstupní sekvencí a vyhledávají tedy globálně optimální řešení. Jsou vhodné především pro vyhledávání kratších motivů a uplatňují se při hledání motivů v eukaryotických organismech, jejichž motivy jsou obecně kratší než u prokaryotických organismů. Metody založené na vyhledávání slov mohou být poměrně rychlé, pokud implementujeme optimalizované datové struktury. Jsou vhodné pro vyhledávání vázaných motivů, které se vyskytují pouze v jedné podobě (nedochází k variacím motivu). Pro typické transkripční motivy, které se objevují ve více podobách (například na určité pozici se mohou se stejnou či různou pravděpodobností vyskytnout dva různé nukleotidy), nejsou výsledky vyhledávání těchto metod tak optimální a často je potřeba je ještě podrobit shlukování. V rámci metod založených na vyhledávání slov také dochází k nacházení většího počtu falešných motivů (metoda oblast sice identifikuje jako motiv, ale ve skutečnosti to motiv není). U pravděpodobnostních metod dochází k odhadu motivů pomocí principů maximální pravděpodobnosti nebo Bayesova odvozování. Pravděpodobnostní přístup k řešení problému zahrnuje také reprezentaci motivu pomocí váhové matice pozicí. Tato matice je obvykle zobrazována jako piktogram, ve kterém je každá pozice reprezentována sloupcem písmen, jejichž výška je shodná s informačním

obsahem této pozice. Výhodou pravděpodobnostních metod je to, že požadují malé množství parametrů vyhledávání a opírají se o pravděpodobnostní modely regulačních oblastí, které jsou velmi citlivé na malé změny ve vstupních datech. Algoritmy založené na pravděpodobnostních metodách jsou často konstruovány pro vyhledávání delších a obecnějších motivů než jsou požadovány pro transkripční motivy. Pravděpodobnostní metody se využívají převážně pro hledání motivů v regulačních oblastech prokaryontních organismů. Výsledky vyhledávání pomocí pravděpodobnostních metod nelze považovat za globálně optimální, ale kvůli pravděpodobnostnímu přístupu pouze lokálně optimální, neboť jsou vstupní data prohledávána pouze lokálně pomocí například Gibbsova vzorkování nebo metody očekávání-maximalizace. [8], [14]

V rámci této práce se budu zabývat pouze algoritmy pracujícími s promotorovými sekvencemi v rámci jednoho genomu. Kvůli obrovskému množství existujících algoritmů v následující kapitole popíšu ty nejznámější a v současné době nejpoužívanější metody.

3.1 Metody založené na vyhledávání slov

Z algoritmů založených na vyhledávání slov je neopomenutelný algoritmus Oligo-Analysis, dále pak Yeast Motif Finder a v současnosti velmi využívaný MDScan.

3.1.1 Oligo-Analysis

Algoritmus Oligo-Analysis je jednoduchou a rychlou metodou izolace motivů ze sekvencí promotorů koregulovaných genů. Koncepčně je jednoduchý a ukázal se jako poměrně efektivní pro vyhledávání již známých motivů ve většině regulačních oblastí kvasinky, dokonce dokázal předpovědět nové zajímavé oblasti. Algoritmus je založen na detekci oligonukleotidů (úsek DNA o 2 a více nukleotidech), které jsou v dané oblasti nadměrně zastoupeny. V rámci algoritmu je definována statistická významnost oligonukleotidu na základě tabulek oligonukleotidových frekvencí, které byly získány ze všech nekódujících sekvencích genomu kvasinky. Rozsah detekce tohoto algoritmu je limitován na poměrně jednoduché vzory, které obsahují krátké motivy s konzervovaným jádrem. [8], [14]

Před použitím tohoto algoritmu je potřeba provést kalibraci. Ta je realizována vložením sekvencí, z nichž chceme vypočítat předpokládané frekvence výskytu oligonukleotidů. Jako sekvence pro výpočet předpokládaných frekvencí se používá většinou soubor všech nekódujících sekvencí genomu studovaného organismu (nejlépe funguje tento algoritmus pro kvasinku). Také je potřeba si zvolit délku oligonukleotidů *w*, které hodláme hledat (například 6 nukleotidů). Je proveden výpočet pozorovaných frekvencí pro každý možný oligonukleotid *b* v rámci všech nekódujících oblastí genomu organismu. Tyto frekvence jsou pak použity pro odhad předpokládaných frekvencí specifických oligonukleotidů $F_e\{b\}$.

Pomocí předpokládaných frekvencí se vypočítává předpokládaný počet výskytů každého oligonukleotidu pomocí vzorce:

$$E(occ\{b\}) = F_e\{b\} \times 2 \times \sum_{i=1}^{S}(L_i - w + 1) = F_e\{b\} * T \quad (3.1)$$

kde *E(occ{b})* je předpokládaný počet výskytů oligonukleotidu *b*, *S* je počet sekvencí použitých pro kalibraci a L_i je délka *i*té sekvence pro kalibraci. Neznámá *T* v rovnici (3.1) reprezentuje celkový počet možných shodných pozic pro vzor délky *w* na obou vláknech sekvencí pro kalibraci. [14]

Po provedené kalibraci vložíme na vstup sekvence nekódujících oblastí, které chceme prozkoumat. Algoritmus spočítá všechny výskyty jednotlivých oligonukleotidů o délce *w* v daných sekvencích. Pravděpodobnost zaznamenání přesně *n* výskytů oligonukleotidu *b* je možné vypočítat pomocí binomické věty:

$$P(occ\{b\} = n) = \frac{T!}{(T-n)! \times n!} \times F_e\{b\}^n \times (1 - F_e\{b\})^{(T-n)} \quad (3.2)$$

Pravděpodobnost zaznamenání *n* nebo více výskytů oligonukleotidu *b* je:

$$P(occ\{b\} \geq n) = \sum_{j=n}^{T} P(occ\{b\} = j) = 1 - \sum_{j=0}^{n-1} P(occ\{b\} = j) \quad (3.3)$$

Vzhledem k tomu, že algoritmus počítá s oběma vlákny a pro každý oligonukleotid je brán v potaz také jeho reverzní komplement, musíme počítat s výskytem palindromických oligonukleotidů, které mají reverzní komplement shodný se svou vlastní sekvencí. Z tohoto důvodu je tedy zaveden počet odlišných oligonukleotidů *D*, který je roven:

$$D = 4^w - (4^w - N_{pal})/2 \quad (3.4)$$

kde *w* je jako v předchozích délka oligonukleotidu a N_{pal} je počet palindromických oligonukleotidů. V případě, že je délka oligonukleotidů *w* sudá, je počet palindromických motivů $N_{pal} = 4^{w/2}$, v ostatních případech je $N_{pal} = 0$. Poslední algoritmem provedený výpočet je výpočet koeficientu významnosti:

$$sig = -log_{10}[P(occ\{b\} \geq n) \times D] \quad (3.5)$$

Koeficient významnosti nabývá největší hodnoty u oligonukleotidu, který se v sekvencích vyskytuje nejčastěji. Obvykle posuzujeme oligonukleotidy s koeficientem významnosti větším než nula. Pokud se na nějakém místě vyskytuje motiv, většinou je zachycen v rámci více než jednoho oligonukleotidu. Konzervované jádro bude mít nejvyšší koeficient významnosti a okolní oligonukleotidy budou mít koeficient nižší. [14]

Algoritmus byl doplněn i o vyhledávání vmezeřených motivů, kdy dochází k výpočtu koeficientu významnosti pro dva trinukleotidy oddělené mezerou. Algoritmus počítá

s délkou mezery od 0 do 16 nukleotidů. Koeficient významnosti je tedy vypočítáván buď na základě kombinace koeficientů významnosti dvou konzervovaných částí vstupních dat nebo na základě předpokládané frekvence konkrétního vmezeřeného motivu vypočtené z kalibračních sekvencí. [8], [9]

Jednoduchost tohoto algoritmu mu nijak neubírá na efektivitě. Velkou nevýhodou je však to, že algoritmus nepovoluje jakékoli variace v rámci oligonukleotidu. [8], [14]

3.1.2 Yeast Motif Finder

Dalším algoritmem založeným na vyhledávání slov je algoritmus Yeast Motif Finder (dále jen YMF). YMF provádí výpočty z-skóre jednotlivých oligonukleotidů a porovnává je. Hodí se převážně pro vyhledávání kratších motivů. Na rozdíl od předchozího algoritmu již počítá s variacemi v rámci motivu. [8]

Vstupem pro YMF je soubor sekvencí promotorů, počet nukleotidů motivu (bez nukleotidů mezery) a matice pravděpodobností přechodu Markovova řetězce řádu *m* sestavená ze všech promotorových sekvencí daného organismu. Nejprve je vypočítán počet výskytů jednotlivých motivů *s*. Podle matice pravděpodobností přechodu je vygenerován soubor náhodných sekvencí DNA stejného počtu a délky jako vstupní sekvence. Tento soubor značíme X a počet výskytů motivu s v souboru X značíme X_s. Dále je vypočítána průměrná hodnota výskytu motivu $E(X_s)$ v rámci souboru. Pro zrychlení algoritmu dochází nejprve k porovnání pravé strany nerovnice (3.6) s nejnižším objeveným z-skóre pro motiv.

$$z_s \leq \frac{N_s - E(X_s)}{\sqrt{E(X_s) - E(X_s)^2}} \tag{3.6}$$

Pokud nerovnice neplatí, pro daný motiv se dále z-skóre nepočítá a dochází ke zkoumání dalšího motivu. Platí-li nerovnice, algoritmus pokračuje výpočtem směrodatné odchylky $\sigma(X_s)$ pomocí vzorce:

$$\sigma(X_s) = \sqrt{E\left({X_s}^2\right) - E(X_s)^2} \tag{3.7}$$

Pomocí průměru a směrodatné odchylky je vypočítáno z-skóre daného motivu s pomocí vzorce:

$$z_s = \frac{N_s - E(X_s)}{\sigma(X_s)} \tag{3.8}$$

Motivy s nejvyšším z-skóre jsou známé transkripční motivy, případně oblasti vhodné k hlubšímu výzkumu. [15]

Tento algoritmus produkuje motivy s nejvyšším z-skóre. Velmi dobře funguje pro sekvence kvasinky, což dokázal při porovnávání s jinými algoritmy, kdy dokázal

detekovat motivy přesněji a na více místech. Za klad algoritmu lze také považovat zahrnutí variací v rámci motivu do vyhledávání. [15]

3.1.3 MDScan

Algoritmus MDScan vybírá několik největších kandidátů na motiv z předcházejícího ChIP-array experimentu. V rámci tohoto experimentu je vybráno několik oblastí nekódující DNA o délce 10 až 1000 nukleotidů, v nichž je předpokládána interakce proteinu s DNA. MDScan pracuje s takovými oblastmi a vytváří modely motivů, které postupně vylepšuje. Model frekvencí pozadí je sestaven pomocí celého genomu organismu a třetího řádu Markovova modelu. [16], [17]

Vstupem pro MDScan je soubor *n* sekvencí DNA vybraných z experimentů ChIP-array a seřazených podle skóre od nejvyššího po nejnižší. MDScan nejprve prozkoumá *t* (~3-20) nejlépe postavených sekvencí souboru. Za předpokladu, že šířka motivu (počet znaků v sekvenci motivu) je rovna *w*, MDScan spočítá každý neopakující se *w*-mer (jádro) objevující se na obou vláknech *t* nejlepších sekvencí a vyhledá všechny *w*-mery v sekvencích *t* s alespoň *m* páry bází shodnými s jádrem (*m*-shody). Počet *m* je definován tak, aby šance, že z páru náhodně generovaných *w*-merů je jeden *m*-shodou druhého, byla menší než 0,15 %. Pro každé jádro jsou vyhledány všechny *m*-shody v sekvencích *t* a použijí se pro vytvoření váhové matice motivu. Dále je pro každý motiv vypočteno skóre pomocí Markovova modelu. Pro vylepšování je ponecháno 10-50 jader s nejvyšším skóre (kandidátů na motivy). [17]

Při procesu vylepšování jsou použity matice jednotlivých ponechaných jader k prozkoumání všech *w*-merů ve zbylých sekvencích. Nový *w*-mer je přidán do váhové matice kandidáta pouze, pokud se tím zvýší skóre matice. V rámci kroku vylepšování dochází k přezkoumání všech částí kandidáta motivu, které jsou již součástí matice motivu. Část je odstraněna z matice, pokud se tím zvýší skóre motivu. Zarovnané části každého motivu se obvykle stabilizují v průběhu deseti vylepšujících iterací. [17]

Z algoritmů založených na vyhledávání slov zmíněných v této práci je MDScan rozhodně nejúčinnější a nejpřesnější. Důvodem je hlavně několikanásobné přezkoumávání již nalezených motivů. Výhodou tohoto algoritmu je jeho dobrá účinnost i při délkách sekvencí nad 300 nukleotidů. [16], [17]

3.2 Pravděpodobnostní metody

Algoritmy založené na pravděpodobnostních metodách využívají převážně dvou těchto metod, očekávání-maximalizace a Gibbsovo vzorkování. Současně mezi hojně užívané algoritmy z této oblasti patří MEME, AlignACE, BioProspector a MotifSampler. [8], [16]

3.2.1 MEME

Algoritmus MEME (Multiple Expectation Maximization Estimation) je založen na metodě očekávání-maximalizace. Vstupem pro tuto metodu je soubor sekvencí DNA a délka motivu. Základem metody očekávání-maximalizace je předpoklad, že každá sekvence vstupního souboru obsahuje příklad motivu. Na počátku je odhadnuta či náhodně vygenerována matice frekvencí nukleotidů na jednotlivých pozicích motivu, označována jako *freq*. Dochází k odhadům (očekávání) pravděpodobnosti $poff_{ij}$, že motiv začíná na pozici *j* v sekvenci *i* vstupního souboru, pomocí Bayesova teorému. Odhadem pro všechny pozice a sekvence vzniká matice *poff*. Matice *poff* je dále využita k přehodnocení (maximalizace) frekvence nukleotidu *l* ve sloupci *c* motivu, tedy matice *freq*, pro každý nukleotid a sloupec (od 1 až po délku motivu). Matice *poff* a *freq* jsou v dalších iteracích vylepšovány, dokud není změna *freq* mezi současnou a předcházející iterací minimální. Pomocí metody očekávání-maximalizace je tedy simultánně objevován model motivu (*freq*) a odhadována pozice počátku daného motivu v sekvencích souboru. Úspěšnost řešení je možné porovnat pomocí pravděpodobnostní funkce, která je závislá na matici *freq*, množství a délce sekvencí v souboru, délce motivu a frekvenci nukleotidů na všech pozicích v sekvenci mimo motiv. Hodnoty pravděpodobnostní funkce pro výstupní modely metody očekávání-maximalizace jsou považovány za lokální maxima této funkce a vyhledává se ten model, jehož hodnota pravděpodobnostní funkce bude nejvyšší, tedy bude globálním maximem. [8], [18]

Algoritmus MEME vylepšil metodu očekávání-maximalizace třemi způsoby. Zaprvé, počáteční matice *freq* je vytvořena na základě subsekvencí trénovacího souboru sekvencí. Zadruhé, ruší předpoklad jedna sekvence=jeden motiv a očekává výskyt žádného až několika motivů v rámci jedné sekvence. Zatřetí, po nalezení motivu je takzvaně pravděpodobnostně vymazán, aby mohl být vyhledáván jiný motiv ve stejném souboru sekvencí. [8], [18]

Algoritmus MEME má v rámci dnes používaných algoritmů poměrně vysokou sensitivitu, úspěšnost nalezení motivu je tedy vyšší než u jiných algoritmů. Algoritmus MEME si dokáže poradit i s rozmanitějšími a delšími vstupními sekvencemi. Velkou výhodou tohoto algoritmu je, že dokáže přizpůsobit šířku motivu a není vázán na jednu určitou. Jeho účinnost tedy není nijak závislá na zvolené šířce motivu. Ukázalo se, že odhady motivů jsou přesnější pro motivy o větší šířce. [16]

3.2.2 AlignACE

Algoritmus AlignACE je založen na metodě Gibbsova vzorkování. Vstupními daty pro Gibbsovo vzorkování jsou soubor N sekvencí DNA a šířka motivu W. V průběhu Gibbsova vzorkování vznikají dvě datové struktury. První z nich je popis vzoru ve formě pravděpodobnostního modelu nukleotidových frekvencí pro každou pozici i od 1 do W a

skládající se z neznámých $q_{i,1}$ až $q_{i,4}$. Tento popis vzoru je ještě doprovázen analogickým pravděpodobnostním popisem frekvencí na pozadí p_1 až p_4, se kterými se nukleotidy vyskytují na pozicích nepopsaných vzorem. Druhou datovou strukturou vznikající v průběhu Gibbsova vzorkování je soubor pozicí a_k, pro k od 1 do N, pro běžný vzor v rámci sekvencí. [8], [19]

Každá iterace Gibbsova vzorkování se skládá ze dvou kroků, prediktivní aktualizace a vzorkování. Prediktivní aktualizace začíná náhodným či specifikovaným výběrem sekvence z ze souboru N sekvencí. Z aktuálních pozic a_k všech sekvencí vyjma z jsou vypočteny frekvence pozadí p_j a popis vzoru $q_{i,j}$. V rámci kroku vzorkování je na každou část sekvence z o šířce W pohlíženo jako na možnou oblast výskytu vzoru. Jsou vypočítány pravděpodobnosti Q_x generování každého úseku x podle současných pravděpodobností vzoru $q_{i,j}$ a také pravděpodobnosti P_x generování tohoto úseku pravděpodobnostmi pozadí p_j. Každému úseku x jsou přiřazeny váhy $A_x=Q_x/P_x$ a z těchto úseků je jeden náhodný vybrán (s pravděpodobností $A_x/\sum jA_j$, kde součet je brán přes všechny úseky). Pozice takto vybraného úseku se stává novým a_z. [8], [19]

Algoritmus AlignACE se od Gibbsova vzorkování liší ve třech hlavních bodech. Jsou prohledávána obě vlákna vstupních sekvencí a překrývající se motivy nejsou povoleny, i kdyby se nacházely každý na jiném vlákně. Současné vyhledávání více motivů je nahrazeno přístupem, kdy dochází k vyhledání vždy jednoho motivu, jeho zamaskování a vyhledávání dalších. AlignACE také využívá vylepšenou více optimální metodu vzorkování. Hlavní statistické skóre využívané algoritmem je takzvané MAP skóre, které udává stupeň nadměrného zastoupení motivu v porovnání s předpokládaným náhodným výskytem motivu v uvažované sekvenci. [8], [16], [20]

Účinnost algoritmu AlignACE je v rámci zde zmíněných algoritmů založených na pravděpodobnostních metodách nejnižší. Oproti očekávání se výstupy algoritmu jeví jako poměrně stabilní při několikanásobném spuštění i přes náhodné výběry v rámci iterací, bohatě tedy stačí spustit algoritmus jednou. Algoritmus funguje poměrně obstojně pro vstupní sekvence o délce do 400 nukleotidů, pro delší sekvence je však poměrně neefektivní. Co se týká šířky motivů, přesněji jsou detekovány motivy o větší šířce. [16]

3.2.3 BioProspector

Základem algoritmu BioProspector je metoda Gibbsova vzorkování. Jedním z rozdílů oproti Gibbsovu vzorkování je využití Markovových modelů nultého až třetího řádu jako charakteristického rozložení nukleotidů na pozadí. Parametry Markovových modelů je možno zadat přímo nebo nechat vypočítat pomocí souboru sekvencí. BioProspector umožňuje vyhledávání palindromických i vmezeřených motivů. Palindromické motivy jsou vyhledávány na obou vláknech sekvence současně a pro vyhledávání vmezeřených motivů

je potřeba zadat předpokládaný rozsah mezery (například 5 až 10 nukleotidů). Kvalita motivu je posuzována pomocí rozložení skóre motivů, které je získáváno pomocí metody Monte Carlo. [8], [21]

V rámci uvedených algoritmů patří BioProspector k těm šikovnějším. Stejně jako AlignACE jsou výstupní data BioProspectoru poměrně stabilní, není třeba algoritmus spouštět vícekrát. Na rozdíl od AlignACE však funguje dobře jak pro kratší tak pro delší vstupní sekvence. Ze zde uvedených algoritmů založených na Gibbsově vzorkování je pro vstupní sekvence nad 300 nukleotidů nejefektivnější. BioProspector je také vhodné použít, pokud vstupní soubor obsahuje velké množství sekvencí. [16]

3.2.4 MotifSampler

Dalším algoritmem založeným na Gibbsově vzorkování je MotifSampler. Metodu Gibbsova vzorkování rozšiřuje ve dvou hlavních prvcích. Prvním je implementace Markovových modelů pro odhad rozložení nukleotidů na pozadí podobně jako u BioProspectoru. Rozdílem mezi BioProspectorem a MotifSamplerem je ten, že MotifSampler využívá vyšší řády Markovových modelů. Druhým rozšířením je přidání odhadu počtu výskytů motivu v každé sekvenci pomocí Bayesova teorému. [8], [22]

Stejně jako AlignACE a BioProspector i MotifSampler je poměrně stabilní a není třeba ho pro stejná vstupní data spouštět vícekrát. MotifSampler je vhodný pro vstupní sekvence kratší délky, neboť pro sekvence o délkách nad 400 nukleotidů se stává poměrně neefektivní podobně jako AlignACE. Algoritmus MotifSampler má ovšem vysokou specificitu, tedy dokáže vcelku přesně určit oblasti neobsahující motiv. [16]

4 VLASTNÍ PROGRAMOVÉ ŘEŠENÍ

4.1 Funkce pro analýzu Oligo-Analysis

K praktickému řešení byl vybrán algoritmus Oligo-Analysis pro jednoduchost jeho zpracování. Pseudokód a vývojový diagram řešení je uveden v příloze A a příloze B.

Pro provedení analýzy sekvencí pomocí algoritmu založeném na principu algoritmu Oligo-Analysis bylo vytvořeno sedm funkcí, konkrétně jsou to funkce *ctifasta*, *komplement*, *kmerpocet*, *kombinace*, *kalibrace*, *ExpOcc* a *OligoAnalysis*. Všechny tyto funkce byly zpracovány v programovacím prostředí Matlab, verze R2012a. Funkce *ctifasta*, *komplement* a *kmerpocet* byly vytvořeny proto, aby bylo možné provést analýzu i v programu Matlab, který neobsahuje bioinformatický toolbox. V bioinformatickém toolboxu jsou tyto funkce k dispozici pod názvy *fastaread*, *seqcomplement* a *nmercount*. Všichni uživatelé programovacího prostředí Matlab však bioinformatický toolbox nemají a právě z tohoto důvodu byly vytvořeny funkce plnící shodnou funkci bez využití tohoto toolboxu. V následujících podkapitolách budou všechny vytvořené funkce popsány.

4.1.1 Funkce pro načtení souboru fasta

Funkce *ctifasta* funguje obdobně jako funkce *fastaread* z bioinformatického toolboxu. Vstupem pro tuto funkci je název souboru fasta, který chceme načíst, v datovém typu *string*. Pomocí funkce *importdata* je načten fasta soubor v datovém typu *structure*. Úkolem funkce *ctifasta* je ze struktury extrahovat názvy a jednotlivé sekvence. Je proto potřeba zjistit, jestli načtený fasta soubor obsahuje pouze jednu sekvenci s názvem či je souborem o více sekvencích. Tento údaj je zjištěn pomocí funkce *length* a uložen do proměnné *pocet*, která je využita v následujícím cyklu. V každém cyklu je do proměnné *hlav* přidán název sekvence a do proměnné *sekv* je přidána odpovídající sekvence. Proměnné *hlav* a *sekv* jsou datového typu *cell* a jsou výstupními proměnnými celé funkce.

4.1.2 Funkce pro tvorbu komplementární sekvence

Funkce *komplement* nahrazuje funkci *seqcomplement* z bioinformatického toolboxu. Slouží k vytvoření komplementární sekvence k sekvenci vložené. Vstupem pro tuto funkci je tedy sekvence v datovém typu *string*, k níž potřebujeme vytvořit sekvenci komplementární. U vložené sekvence nezáleží na tom, zda jsou nukleotidy psány velkým či malým písmem. Toto je ošetřeno na začátku funkce, kde dochází k přepisu všech malých písmen na velká. Dále v rámci cyklu dochází k procházení sekvence po jednotlivých nukleotidech a pomocí přepínače jsou zapisovány nukleotidy komplementární do proměnné *kompl*. Tato proměnná

je výstupní proměnnou funkce *komplement*. V případě, že se ve vstupní sekvencí objeví jiný znaky než A, T, G a C, reprezentující nukleotidy s bázemi adenin (A), thymin (T), guanin (G) a cytosin (C), dojde pouze k přepsání tohoto znaku do komplementární sekvence.

4.1.3 Funkce pro zjištění počtu oligonukleotidů

Funkce *kmerpocet* vykonává stejnou funkci jako *nmercount* z bioinformatického toolboxu. Funkce slouží k zjištění počtu všech oligonukleotidů o délce *k* nacházejících se v prohledávané sekvenci. Pro správné fungování této funkce musí být tedy na vstup vložena sekvence, v níž chceme oligonukleotidy spočítat, a délka vyhledávaných oligonukleotidů formou čísla. Pomocí cyklu funkce prochází celou sekvenci po oligonukleotidech o délce *k*. Nejprve je zjištěno, jestli se ten konkrétní oligonukleotid v sekvenci už vyskytl. V rámci podmínky je oligonukleotid zařazen jako nový v případě, že se v již prozkoumané části sekvence ještě neobjevil, nebo je k počtu shodných, dříve nalezených, oligonukleotidů připočtena jednička. Takovýmto způsobem v cyklu vzniká výstupní proměnná datového typu *cell*, která obsahuje jednotlivé nalezené oligonukleotidy v prvním sloupci a jejich počty ve sloupci druhém.

4.1.4 Funkce pro vypsání všech oligonukleotidů

Funkce *kombinace* je vhodná k předpřipravení proměnné k dalšímu zápisu hodnot. V dalších funkcích je potřeba vytvořit proměnnou, která v prvním sloupci obsahuje všechny možné oligonukleotidy o délce *k* a do druhého sloupce je možné zapisovat hodnoty. Právě k tomu účelu slouží funkce *kombinace*. Na vstup je nutné přivést pouze hodnotu *kmer*, což je číslo, které udává délku zkoumaných oligonukleotidů. Je vytvořena *kmer*-rozměrná matice indexů. Kombinací indexů vznikají všechny možné kombinace čísel 1 až 4 pro délku *kmer*. Převedením indexů na nukleotidy tak, že 1 reprezentuje A, 2 C, 3 T a 4 G, jsou získány všechny možné kombinace nukleotidů. Tyto oligonukleotidy jsou vloženy do prvního sloupce výstupní proměnné datového typu *cell*. Druhý sloupec výstupní proměnné obsahuje nuly.

4.1.5 Funkce pro kalibraci

Pomocí funkce *kalibrace* získá uživatel kalibrační matici potřebnou k analýze Oligo-Analysis. K provedení kalibrace je třeba znát kalibrační sekvence a délku zkoumaných oligonukleotidů. Vstupem pro funkci *kalibrace* je tedy název fasta souboru, který obsahuje všechny kalibrační sekvence, a číslo udávající délku oligonukleotidů. Kalibrace pro algoritmus Oligo-Analysis by měla být provedena z nekódujících úseků sekvence zkoumaného organismu, jak je uvedeno v teoretické části práce. Z toho důvodu je funkce pro kalibraci naprogramována pro fasta soubor, který neobsahuje pouze jednu ale více

sekvencí. Kalibrovat pomocí jedné jediné sekvence postrádá smysl, neboť by výsledek kalibrace byl velmi nevěrohodný. Po načtení kalibračních sekvencí je pomocí výše zmíněné funkce *kombinace* vytvořena proměnná *komb*. Do této proměnné jsou v následujícím cyklu připočítávány výskyty jednotlivých oligonukleotidů v kalibračních sekvencích. Cyklus probíhá pro každou sekvenci ze souboru v případě, že splňuje podmínku, která povoluje pouze sekvence o délce 100 nukleotidů a více. Hodnota byla zvolena proto, že sekvence dlouhá sto nukleotidů je již dostatečně reprezentativním vzorkem nekódující oblasti, kratší sekvence vnášejí do kalibrace nepřesnosti. Sekvence splňující podmínku je podrobena výše uvedené funkci *kmerpocet*, která zjistí výskyty jednotlivých oligonukleotidů a tyto výskyty jsou připočteny k již zaznamenaným výskytům v rámci proměnné *komb*. Po prozkoumání všech sekvencí je každý počet výskytů daného oligonukleotidu vydělen celkovým počtem všech nalezených oligonukleotidů v kalibračních sekvencích, čímž získáme očekávanou frekvenci tohoto oligonukleotidu v nekódujících oblastech daného organismu. Výstupem funkce je tedy proměnná datového typu *cell*, která má v prvním sloupci buněk uloženy jednotlivé oligonukleotidy a v druhém sloupci k nim příslušné očekávané frekvence.

4.1.6 Funkce pro zjištění očekávaného výskytu

Na funkci *ExpOcc* je možné nahlížet jako na doplňkovou funkci. Její použití v rámci analýzy Oligo-Analysis není nutné. Pomocí této funkce získá uživatel představu o tom, jak by složení zkoumané sekvence mělo vypadat v případě, že se v něm nevyskytují žádné motivy a jde o typickou nekódující sekvenci daného organismu. Na vstup funkce je nutné přivést název fasta souboru sekvence či sekvencí, pro které požaduje uživatel zjistit očekávané výskyty oligonukleotidů, a název kalibrační matice vytvořené pro odpovídající délku oligonukleotidů. V rámci funkce je vypočteno, na kolik oligonukleotidů o dané délce je možné sekvence rozdělit. Tímto číslem jsou vynásobeny jednotlivé očekávané frekvence pro každý oligonukleotid. Výstupem je tedy proměnná datového typu *cell*, která v prvním sloupci obsahuje jednotlivé oligonukleotidy a druhý sloupec udává, kolikrát by se měl daný oligonukleotid v sekvenci vyskytnout.

4.1.7 Funkce pro analýzu OligoAnalysis

Pomocí funkce *OligoAnalysis* je možné provést celou analýzu Oligo-Analysis. Tato funkce využívá výše popsané funkce *ctifasta, kombinace, kmerpocet* a také kalibrační matici získanou pomocí funkce *kalibrace*. Vstupními daty pro tuto funkci jsou tedy název fasta souboru sekvence nebo sekvencí, ve kterých uživatel vyhledává motivy, a název souboru kalibrační matice podle toho, jaká délka oligonukleotidů má být zkoumána.

Po načtení vstupních dat je potřeba zjistit počet výskytů jednotlivých oligonukleotidů v rámci všech sekvencí na templátovém i komplementárním vlákně. První podmínka

rozděluje tento výpočet na případy, že byla vložena pouze jedna vstupní sekvence nebo se ve vstupním fasta souboru nacházelo vstupních sekvencí více. V prvním případě je výpočet jednoduchý a používá pouze funkci *kmerpocet*, která zjistí výskyty jednotlivých oligonukleotidů a ty se pak přiřadí k oligonukleotidům v přehlednější proměnné. Následně ten samý proces probíhá pro komplementární vlákno. Je vytvořen reverzní komplement originální sekvence a je vypočítán celkový počet pozic o dané délce na vstupní sekvenci a jejím komplementu. Ve druhém případě je potřeba uvedené výpočty provést pro každou vstupní sekvenci, což zajišťuje v podmínce vnořený cyklus.

Dalším krokem je výpočet pravděpodobnosti *N* a více výskytů jednotlivých oligonukleotidů využívající vzorce (3.2) a (3.3) z teoretické části práce. Cyklus prochází jednotlivé oligonukleotidy, zjišťuje jejich reálný výskyt ve vstupních sekvencích, v rámci vnořeného cyklu počítá pravděpodobnost *N-1* a méně výskytů daného oligonukleotidu a nakonec tuto pravděpodobnost odečte od jedné, čímž je získána pravděpodobnost *N* a více výskytů. Pro správné fungování výpočtu pravděpodobnosti *N-1* a méně výskytů bylo potřeba vzorec (3.2) upravit pro případy, kdy *n* je nula a jedna. V případě, že *n* je 0, po dosazení do vzorce získáme:

$$P(occ\{b\} = 0) = \frac{T!}{(T-0)! \times 0!} \times (F_e\{b\})^0 \times (1 - F_e\{b\})^{(T-0)} \tag{4.1}$$

a po úpravě tedy:

$$P(occ\{b\} = 0) = (1 - F_e\{b\})^T \tag{4.2}$$

V případě, že *n* je 1, po dosazení do vzorce získáme:

$$P(occ\{b\} = 1) = \frac{T!}{(T-1)! \times 1!} \times (F_e\{b\})^1 \times (1 - F_e\{b\})^{(T-1)} \tag{4.3}$$

a po úpravě:

$$P(occ\{b\} = 1) = T \times F_e\{b\} \times (1 - F_e\{b\})^{(T-1)} \tag{4.4}$$

Pro všechny ostatní případy je pravděpodobnost vypočítána pomocí vzorce (3.2) s upraveným prvním zlomkem na tvar:

$$\frac{T!}{(T-n)! \times n!} = \frac{T \times (T-1) \times \ldots \times (T-n+1)}{n!} \tag{4.5}$$

Čitatel zlomku je vypočítán pomocí funkce *prod*, která vynásobí všechny prvky předem připraveného vektoru hodnot. Při výpočtu pravděpodobnosti konkrétního výskytu oligonukleotidu nedokáže prostředí Matlab pracovat s přesným číslem očekávané frekvence oligonukleotidu (ve vzorci $F_e\{b\}$), dochází k zaokrouhlování této frekvence na 5 desetinných míst, což vnáší do výpočtu chybu. Nejzřetelnější je chyba v případě, že celková pravděpodobnost *N-1* a méně výskytů se blíží jedné, kdy může v důsledku chyby dojít k tomu, že celková suma pravděpodobností překročí hodnotu 1. Překročení hodnoty 1 by

znamenalo nemožnost dalších výpočtů pro ten daný nukleotid, proto je zavedena podmínka, která v případě překročení jedničky hodnotu přepíše na 0,999999999999999.

V následující části funkce je zjištěn počet odlišných oligonukleotidů pomocí vzorce (3.4). V podmínce se mění parametr výpočtu v závislosti na tom, zda je délka zkoumaných oligonukleotidů sudé či liché číslo.

Předposlední část funkce zajišťuje výpočet koeficientu významnosti nalezených výskytů jednotlivých oligonukleotidů podle vzorce (3.5). V rámci cyklu je vypočítáván koeficient významnosti pro každý oligonukleotid.

Nakonec jsou uloženy informace o oligonukleotidech s kladným koeficientem významnosti do výstupní proměnné. První sloupec výstupní proměnné datového typu *cell* obsahuje oligonukleotidy, v druhém sloupci k nim příslušné koeficienty významnosti. Třetí sloupec obsahuje očekávaný počet výskytu daného oligonukleotidu, který je možné porovnat s nalezeným počtem výskytů uvedeným ve čtvrtém sloupci výstupní proměnné.

4.2 Provedení kalibrace

Pro správné fungování algoritmu je potřeba před spuštěním funkce *OligoAnalysis* provést kalibraci pro zkoumaný organismus. Je nutné dbát na správný výběr kalibračních sekvencí. Nevhodně zvolené kalibrační sekvence mohou vést k nesprávné matici předpokládaných frekvencí oligonukleotidů, jejíž použití během analýzy vede k nepřesným výsledkům. Správně zvolené kalibrační sekvence by měly obsahovat veškeré nekódující oblasti genomu zkoumaného organismu.

Tato práce se dále zabývá účinností algoritmu pro vyhledávání motivů v částech genomu kvasinky *Saccharomyces cerevisiae*. Pro veškeré zde uvedené analýzy byla proto využita data získaná kalibrací pro tuto kvasinku. Jako kalibrační sekvence bylo potřeba použít soubor nekódujících sekvencí kvasinky *Saccharomyces cerevisiae*, které byly získány z databáze Saccharomyces Genome Database dostupné na http://yeastgenome.org/. K souboru dat je možné se na této stránce dostat přes záložky Download, dále Sequence a zde je nalezneme v rámci S288C Reference Genome ve složce intergenic. Složka obsahuje fasta soubor s názvem NotFeature. V tomto souboru se nachází 6692 nekódujících sekvencí o různé délce, celková délka všech sekvencí v sumě je 2 911 459 bp (párů bází).

Aby nebylo nutné před každou analýzou v rámci této práce provádět kalibraci, byla realizována kalibrace s kalibračními sekvencemi kvasinky pro délku oligonukleotidů 4 až 8 a výstupy jednotlivých kalibrací (kalibrační matice očekávaných frekvencí oligonukleotidů) byly uloženy. Tyto matice s názvy komb4.mat až komb8.mat jsou k dispozici v rámci elektronických výstupů práce.

4.3 Validace algoritmu na umělých sekvencích

Pro ověření správného fungování vytvořených funkcí byla provedena validace na umělých sekvencích s uměle vloženými motivy. K tomu účelu byl vytvořen soubor sekvencí umela.fasta, který obsahuje dvě sekvence s náhodným rozložením nukleotidů o délce 200 nukleotidů. Do každé sekvence byl uměle vložen jeden motiv na jedenácti pozicích po celé délce sekvence. Vzor motivu z první sekvence je TAATCCGA a z druhé sekvence TAATATTA. První vzor motivu byl zvolen čistě náhodně, druhý byl sestaven tak, aby bylo s jeho pomocí možné posoudit schopnost algoritmu odhalit palindromické motivy. Sekvence byly vkládány na vstup algoritmu postupně společně se všemi pěti kalibračními maticemi, tedy byly vyhledávány motivy o délce čtyř až osmi nukleotidů.

V rámci analýzy oligonukleotidů délky 4 byly nalezeny všechny části předem uměle vložených motivů. Tabulka 4.1 zobrazuje výsledky analýzy pro oligonukleotidy, které patří mezi subsekvence originálních motivů. Nejvyšší koeficienty významnosti, vyšší než 11, byly vypočítány pro motiv TAAT a k němu komplementární ATTA. Tyto dva oligonukleotidy se v obou sekvencích vyskytovaly nejvíce, protože jsou součástí obou vnesených motivů a u palindromického motivu jej najdeme i v komplementární sekvenci. Vložených motivů bylo celkem 11 od každého, tedy díky palindromu bychom očekávali reálný výskyt zmíněných dvou oligonukleotidů 33. Dva další nalezené motivy se pravděpodobně v sekvencích vyskytly náhodně. Koeficientu významnosti přibližně 6 dosahovaly motivy ATCC, TCCG a CCGA a k nim komplementární TAGG, AGGC a GGCT. Nejnižší koeficient významnosti byl zaznamenán u motivu AATC a jeho komplementu TTAG přesto, že se reálně v sekvencích vyskytoval stejně často jako předchozí oligonukleotidy. Důvodem je vyšší očekávaný výskyt tohoto oligonukleotidu oproti předchozím. Co se druhého motivu týče, přestože se jednotlivé oligonukleotidy vyskytují častěji, jsou koeficienty významnosti pro motivy AATA,TATT a ATAT nižší. To je způsobeno velkou četností těchto oligonukleotidů v nekódujících oblastech kvasinky.

Tabulka 4.1: Výsledky vyhledávání motivů o délce 4 v uměle vytvořených sekvencích

Motiv	**Koef. význ.**	**Očekávaný výskyt**	**Reálný výskyt**	**Kompl. motiv**	**Koef. význ.**	**Očekávaný výskyt**	**Reálný výskyt**
TAAT	11,41	7,16	35	**ATTA**	11,42	7,12	35
AATC	1,53	3,40	12	**GATT**	1,58	3,36	12
ATCC	3,19	1,89	11	**GGAT**	3,24	1,87	11
TCCG	6,91	1,26	13	**CGGA**	6,80	1,28	13
CCGA	6,63	1,08	12	**TCGG**	6,51	1,10	12
AATA	2,95	9,65	26	**TATT**	3,07	9,50	26
ATAT	1,32	10,87	24				

Analýzou provedenou pro délku oligonukleotidů 5 bylo také dosaženo obstojných výsledků (Tabulka 4.2). Co se prvního motivu týče, nejvyšší koeficienty významnosti dosahovaly oligonukleotidy z druhé části motivu (ATCCG, TCCGA a jejich komplementy). Koeficienty významnosti oligonukleotidů TAATC, AATCC a jejich komplementů se pohybovaly kolem 7, kvůli vyššímu očekávanému výskytu než u předchozích. Palindromický motiv byl rozeznán velmi dobře, všem jeho subsekvencím byl přiřazen koeficient významnosti více než 9.

Tabulka 4.2: Výsledky vyhledávání motivů o délce 5 v uměle vytvořených sekvencích

Motiv	**Koef. význ.**	**Očekávaný výskyt**	**Reálný výskyt**	**Kompl. motiv**	**Koef. význ.**	**Očekávaný výskyt**	**Reálný výskyt**
TAATC	6,83	0,91	12	**GATTA**	6,74	0,93	12
AATCC	7,95	0,56	11	**GGATT**	8,16	0,53	11
ATCCG	10,52	0,30	11	**CGGAT**	10,35	0,32	11
TCCGA	10,35	0,33	11	**TCGGA**	10,40	0,32	11
TAATA	9,78	2,79	22	**TATTA**	9,92	2,72	22
AATAT	9,48	3,21	23	**ATATT**	9,55	3,19	23

V následující tabulce (Tabulka 4.3) jsou uvedeny výsledky vyhledávání hexanukleotidů. Všechny subsekvence vložených motivů vykazují koeficient významnosti větší než 10, což potvrzuje dobré fungování vyhledávacího algoritmu. Při této délce vyhledávaných oligonukleotidů získáváme poprvé shodný počet nalezených částí motivů s počtem vložených motivů. Subsekvence prvního motivu byly nalezeny 11 krát a subsekvence palindromu 22 krát. Při délce oligonukleotidů 6 už tedy není pravděpodobné, že se oligonukleotid ve zkoumaných sekvencích vyskytuje vícekrát pouze náhodně.

Tabulka 4.3: Výsledky vyhledávání motivů o délce 6 v uměle vytvořených sekvencích

Motiv	**Koef. význ.**	**Očekávaný výskyt**	**Reálný výskyt**	**Kompl. motiv**	**Koef. význ.**	**Očekávaný výskyt**	**Reálný výskyt**
TAATCC	10,26	0,14	11	**GGATTA**	10,24	0,13	11
AATCCG	10,12	0,08	11	**CGGATT**	11,68	0,09	11
ATCCGA	11,08	0,08	11	**TCGGAT**	11,68	0,08	11
TAATAT	10,19	1,01	22	**ATATTA**	11,68	1,01	22
AATATT	11,68	1,07	22				

Podobných výsledků bylo dosaženo při vyhledávání motivů o délce 7 (Tabulka 4.4). Všechny subsekvence byly nalezeny ve správném počtu a jejich koeficienty významnosti se

pohybovaly okolo 10. Nepatrně nižší koeficient významnosti (méně než 10) najdeme u motivů AATCCGA a CGGATTA, kvůli vyššímu očekávanému výskytu.

Tabulka 4.4: Výsledky vyhledávání motivů o délce 7 v uměle vytvořených sekvencích

Motiv	**Koef. význ.**	**Očeк. výskyt**	**Reálný výskyt**	**Kompl. motiv**	**Koef. význ.**	**Očeк. výskyt**	**Reálný výskyt**
TAATCCG	11,09	0,02	11	**CGGATTA**	9,73	0,03	11
AATCCGA	9,75	0,02	11	**TCGGATT**	11,09	0,03	11
TAATATT	11,09	0,35	22	**AATATTA**	11,09	0,36	22

V rámci vyhledávání pro délku 8 byly nalezeny uměle vložené oligonukleotidy v přesném počtu (Tabulka 4.5). U motivu TAATCCGA bylo dosaženo koeficientu významnosti 8,95, což je méně než pro jeho komplement. Důvodem je opět nepatrně větší očekávaný výskyt motivu. Palindromický motiv byl spolehlivě rozeznán a ohodnocen koeficientem významnosti 10,48.

Tabulka 4.5: Výsledky vyhledávání motivů o délce 8 v uměle vytvořených sekvencích

Motiv	**Koef. význ.**	**Očeк. výskyt**	**Reálný výskyt**	**Kompl. motiv**	**Koef. význ.**	**Očeк. výskyt**	**Reálný výskyt**
TAATCCGA	8,95	0,01	11	**TCGGATTA**	10,48	0,01	11
TAATATTA	10,48	0,14	22				

Provedením validace algoritmu na umělých sekvencích byla zjištěna správná funkce algoritmu. Algoritmus je schopen nalézt uměle vnesené motivy, včetně palindromů. Vhodnější je vyhledávat motivy o větší délce, protože je jejich nalezení snadnější, a když je určitý oligonukleotid označen za významný, není pravděpodobné, že by šlo o náhodu. V rámci této kapitoly jsou uvedeny pouze výsledky týkající se vložených motivů. Pomocí analýzy bylo nalezeno i pár jiných oligonukleotidů s kladným koeficientem významnosti, ale tyto nejsou z hlediska validace zajímavé. Veškeré získané výsledky jsou uvedeny v elektronické příloze.

4.4 Analýza rodin MET a PDR u kvasinky *S. cerevisiae*

K ověření funkčnosti algoritmu pro konkrétní organismus byla provedena podobná analýza, jakou uvádějí autoři metody J. van Helden a kolektiv ve svém článku [14]. V rámci analýzy autorů jsou motivy vyhledávány vždy v celé rodině koregulovaných genů kvasinky. Z důvodu obsáhlosti analýz jednotlivých rodin byly pro tuto práci zvoleny pouze dvě rodiny, které byly analyzovány, a výsledky analýzy porovnány s výsledky autorů.

První rodinou koregulovaných genů je rodina MET, která zahrnuje geny, jejichž transkripce je potlačována methioninem. Z této rodiny byly do analýzy zařazeny regulační oblasti genů MET3, MET2, MET14, MET6, SAM1, SAM2, MET1, MET30 a MUP3. Pro každý z těchto genů byla z databáze NCBI extrahována oblast 500 nukleotidů před začátkem ORF daného genu. Tyto sekvence byly spojeny do jednoho fasta souboru s názvem MET_family.

Druhou zkoumanou rodinou koregulovaných genů je rodina PDR zahrnující geny, které zajišťují rezistenci vůči lékům reagujícím pozitivně na více než jednu nemoc. Pro analýzu této rodiny byly použity regulační oblasti genů YOR1, PDR11, PDR10, GAS1, STE6, SNQ2 a PDR5. I pro tyto geny byly extrahovány z databáze NCBI oblasti 500 nukleotidů před začátkem ORF a tyto sekvence spojeny do jednoho fasta souboru s názvem PDR_family.

V rámci následujících kapitol budou vyhodnoceny výsledky pro oligonukleotidy porovnatelné s výsledky autorů metody. Celkové výsledky analýzy jsou uvedeny v elektronické příloze.

4.4.1 Rodina MET

Výsledky analýzy rodiny MET autorů metody poukazují na dva motivy vyskytující se v této rodině. Jsou to motivy GTCACGTG a AAACTGTGG. Totožné motivy byly nalezeny i pomocí vytvořené funkce *OligoAnalysis*. Analýza rodiny MET byla provedena pro všech pět různých délek oligonukleotidů, tedy 4 až 8.

V rámci analýzy pro délku oligonukleotidů 4 (Tabulka 4.6) byly nalezeny oba výše zmíněné motivy. První motiv byl rozpoznán v rámci pěti na sebe navazujících oligonukleotidů, z nichž nejvyšší koeficient významnosti dosahoval oligonukleotid CACG. Autoři metody uvádí vždy pouze nejvyšší koeficient významnosti v rámci motivu a v tomto případě je téměř totožný s výsledkem funkce *OligoAnalysis*. Tento motiv je palindromem, tedy zahrnuje v sobě i svou komplementární variantu. Co se druhého motivu týče, byla nalezena pouze jeho středová část, konkrétně oligonukleotid TGTG. Funkce *OligoAnalysis* tomuto oligonukleotidu přiřadila koeficient významnosti 3,25, což je o téměř 2 více než hodnota uváděná autory. Pro tento motiv byla nalezena i část jeho komplementu, konkrétně oligonukleotidy CCAC a CACA. Poměrně vysoký koeficient významnosti vykazoval také oligonukleotid TGCA, který však nezapadá do žádného autory zmíněného motivu. Mohlo dojít k čistě náhodnému vícenásobnému výskytu tohoto oligonukleotidu nebo může být vyšší významnost způsobena odlišnými kalibračními sekvencemi a tedy i odlišnými očekávanými frekvencemi jednotlivých oligonukleotidů.

Tabulka 4.6:Výsledky vyhledávání motivů délky 4 v rodině MET

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
TCAC	1,43	27,65	48	4,4
CACG	4,50	13,99	37	
ACGT	1,30	23,54	42	
CGTG	4,31	14,25	37	
GTGA	1,42	27,66	48	
TGTG	3,25	30,14	58	1,5
CCAC	0,02	16,68	28	-
CACA	3,05	30,62	58	
TGCA	2,26	34,22	60	-

Vyhledáváním motivů délky 5 bylo také dosaženo podobných výsledků, jako uvádí autoři (Tabulka 4.7). Palindromický motiv byl rozpoznán v rámci čtyř navazujících oligonukleotidů, s nejvyšším dosaženým koeficientem významnosti 3,66. Tato hodnota není příliš odlišná od autory uváděného koeficientu významnosti 4,1. Rozdíl hodnot mohl být způsoben odlišnými kalibračními sekvencemi, tedy odlišnými očekávanými frekvencemi, nebo vyšší četností výskytu některého z oligonukleotidů. Z druhého motivu byla opět nalezena pouze středová část TGTGG a její komplement, kdy oba tyto oligonukleotidy vykazují hodnotu koeficientu významnosti blízkou 2,3 udávanou autory. Koeficient významnosti větší než jedna byl v rámci analýzy přiřazen ještě oligonukleotidu CACAC a jeho komplementu GTGTG. Tato hodnota koeficientu významnosti ještě není natolik vysoká, aby se dalo říci, že jde pravděpodobně o motiv. Stále je možné, že jde pouze o náhodný vícenásobný výskyt tohoto oligonukleotidu.

Analýza oligonukleotidů o délce 6 je jediná, pro kterou autoři uvádí konkrétnější výsledky k jednotlivým oligonukleotidům. V posledním sloupci tabulky (Tabulka 4.8) jsou tedy vypsány koeficienty významnosti příslušné jednotlivým oligonukleotidům na daném řádku. Palindromický motiv byl opět odhalen s vysokou přesností, dokonce by se ze získaných dat dalo motiv rozvést na konci o další pravděpodobně následující nukleotidy. Druhý motiv sice nalezen byl, ale jeho koeficienty významnosti jsou poměrně nízké hodnoty. Lepších hodnot dosahují oligonukleotidy komplementu motivu, kvůli nižšímu očekávanému výskytu těchto oligonukleotidů.

Tabulka 4.7: Výsledky vyhledávání motivů délky 5 v rodině MET

Motiv	Výsledky funkce *OligoAnalysis*			Koeficient významnosti uváděný autory [14]
	Koeficient významnosti	Očekávaný výskyt	Reálný výskyt	
TCACG	2,87	4,27	17	4,1
CACGT	3,66	4,61	19	
ACGTG	3,61	4,64	19	
CGTGA	2,98	4,18	17	
TGTGG	2,29	5,24	18	2,3
CCACA	2,35	5,19	18	-
CACAC	1,39	6,17	18	-
GTGTG	1,51	6,04	18	-

Tabulka 4.8: Výsledky vyhledávání motivů délky 6 v rodině MET

Motiv	Výsledky funkce *OligoAnalysis*			Koeficient významnosti uváděný autory [14]
	Koeficient významnosti	Očekávaný výskyt	Reálný výskyt	
GTCACG	0,60	0,74	6	0,7
TCACGT	5,30	1,63	14	6,1
CACGTG	9,44	1,10	18	7,0
ACGTGA	5,81	1,49	14	-
CGTGAC	0,67	0,72	6	-
AACTGT	0,96	2,91	12	0,9
ACTGTG	0,25	1,61	8	0,6
TGTGGC	0,97	1,26	8	0,5
GCCACA	1,25	1,14	8	-
CACAGT	0,14	1,67	8	-
ACAGTT	1,20	2,74	12	-

Vyhledáváním motivů délky 7 byly získány uspokojivé výsledky (Tabulka 4.9). Palindromický motiv byl rozpoznán bez problému a středová část byla ohodnocena koeficientem významnosti vyšším než 8, což odpovídá hodnotě uváděné autory metody. Lepších hodnot než v předchozích analýzách dosáhl druhý motiv a jeho komplement. Funkce *OligoAnalysis* oligonukleotidům odpovídajícím druhému motivu přiřadila koeficient

významnosti vyšší než 2. Autoři uvádí hodnotu 4,8. Rozdíl hodnot může být způsoben odlišnými kalibračními sekvencemi nebo vyšší četností výskytu daných oligonukleotidů v sekvencích zkoumaných autory. Rozdílné hodnoty však nemění fakt, že byl motiv rozpoznán.

Tabulka 4.9: Výsledky vyhledávání motivů délky 7 v rodině MET

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
GTCACGT	1,64	0,38	6	8,2
TCACGTG	8,46	0,52	14	
CACGTGA	8,68	0,47	14	
ACGTGAC	2,27	0,29	6	
AACTGTG	2,30	0,55	8	4,8
ACTGTGG	2,14	0,31	6	
CCACAGT	1,76	0,36	6	-
CACAGTT	2,82	0,58	8	

Analýzou motivů o délce 8 (Tabulka 4.10) byly získány nepatrně nižší koeficienty významnosti pro palindromický motiv než v předchozích. Bylo však dosaženo vyšší hodnoty, než jakou uvádí autoři metody. Při této délce vyhledávaných oligonukleotidů však dosahoval nejlepších hodnot druhý motiv. Byly nalezeny všechny jeho části a ta nejkonzervovanější dosáhla koeficientu významnosti většího než 4, tedy můžeme předpokládat náhodný výskyt tohoto oligonukleotidu alespoň jednou v rámci 10 000 rodin. Oligonukleotid ACTGTGGC má sice nižší koeficient významnosti, ale může se jednat o prodloužení motivu, které autoři metody ve své analýze neuvádějí.

Tabulka 4.10: Výsledky vyhledávání motivů délky 8 v rodině MET

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
GTCACGTG	2,65	0,20	6	3,2
TCACGTGA	7,79	0,22	10	
CACGTGAC	3,74	0,13	6	
AAACTGTG	3,64	0,24	7	5,2
AACTGTGG	4,15	0,11	6	
ACTGTGGC	0,90	0,10	4	
GCCACAGT	1,01	0,09	4	-
CCACAGTT	3,13	0,16	6	
CACAGTTT	3,21	0,28	7	

4.4.2 Rodina PDR

V rodině koregulovaných genů PDR J. van Helden a kolektiv poukazují opět na dva nalezené motivy. Jde o motivy TCCGTGGA a TCCGCGGA lišící se pouze středovým nukleotidem, který byl pravděpodobně pozměněn mutací, ale to nemění funkčnost motivu. Motiv s C uprostřed je na první pohled palindromem. I pro tuto rodinu byly vyhledávány motivy pro délky od 4 do 8.

Z výsledků vyhledávání motivů o délce 4 nukleotidů je patrné nalezení obou zmíněných motivů (Tabulka 4.11). Prvnímu motivu byly v rámci této konkrétní analýzy přiřazeny poměrně nízké hodnoty koeficientu významnosti okolo 1. Tyto hodnoty však korespondují s hodnotou 1,5 uváděnou autory metody. Komplement prvního motivu dosahoval podobných hodnot jako motiv samotný. Oproti tomu oligonukleotidy tvořící střed palindromického motivu dosahovaly koeficientu významnosti většího než 4. Okrajové části byly sice také rozpoznány, ale hodnota koeficientu významnosti u nich nepřesáhla jedničku. Autoři pro tento motiv v rámci analýzy oligonukleotidů o délce 4 uvádí nejvyšší dosaženou hodnotu koeficientu významnosti 6,9. Rozdíl hodnot byl opět pravděpodobně způsoben rozdílnými kalibračními sekvencemi či odlišnou četností výskytu některého z oligonukleotidů v sekvencích zkoumaných autory metody.

Tabulka 4.11: Výsledky vyhledávání motivů délky 4 v rodině PDR

Motiv	Výsledky funkce *OligoAnalysis*			Koeficient významnosti uváděný autory [14]
	Koeficient významnosti	Očekávaný výskyt	Reálný výskyt	
TCCG	0,80	11,11	23	1,5
CCGT	0,36	11,36	22	
CGTG	1,16	11,09	24	
GTGG	1,15	13,13	27	
CCAC	1,23	12,98	27	-
CACG	1,27	10,88	24	
ACGG	0,48	11,08	22	
CGGA	0,69	11,33	23	
TCCG	0,80	11,11	23	6,9
CCGC	4,63	9,29	29	
CGCG	4,24	8,09	26	
GCGG	4,33	9,62	29	
CGGA	0,69	11,33	23	

V rámci analýzy délky 5 (Tabulka 4.12) byly získány nepatrně lepší výsledky pro první motiv. Nejvyšší hodnota koeficientu významnosti pro tento motiv přesáhla hodnotu 2. Opět je hodnota podobná hodnotě uváděné autory. Komplementární oligonukleotidy k prvnímu motivu vykazovaly koeficient významnosti téměř 2. Palindromický motiv byl spolehlivě rozeznán s nejvyšší hodnotou koeficientu významnosti větší než 5. Odchylka od hodnoty uváděné autory není tolik důležitá jako fakt, že právě při analýze délky 5 bylo pro tento motiv dosaženo nejvyšší hodnoty ze všech analýz, a to konkrétně pro oligonukleotid CCGCG.

Výsledky vyhledávání motivů délky 6 potvrzují výskyt obou motivů (Tabulka 4.13). Hodnota koeficientu významnosti prvního motivu by v případě této analýzy podle autorů měla dosahovat nejvyšších hodnot, což ovšem výsledky funkce *OligoAnalysis* nepotvrzují. Na rozdíl od autory udávané hodnoty vyšší než 7, je pomocí funkce dosažena hodnota pouze nepatrně větší než 3, která je v rámci jedné z následujících analýz překročena. Tato odlišnost může být způsobena vyšší četností motivu ve vzdálenější části sekvencí od kódujících oblastí použitých autory, které v rámci funkce nebyly zkoumány. Oligonukleotidům komplementárním k prvnímu motivu byly přiznány hodnoty okolo dvou. Byl také nalezen oligonukleotid CACGGA navazující na předchozí oligonukleotidy, který se v klasickém

motivu neobjevuje. Středová část palindromického motivu byla detekována lépe než autory. Okraje motivu byly detekovány, ale s podstatně menším koeficientem významnosti.

Tabulka 4.12: Výsledky vyhledávání motivů délky 5 v rodině PDR

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
CCGTG	1,62	2,00	10	3,3
CGTGG	2,09	2,15	11	
CCACG	1,99	2,20	11	-
CACGG	1,80	1,90	10	
TCCGC	1,31	2,63	11	7,1
CCGCG	5,10	1,59	13	
CGCGG	4,82	1,69	13	
GCGGA	1,15	2,74	11	

Tabulka 4.13: Výsledky vyhledávání motivů délky 6 v rodině PDR

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
CCGTGG	3,03	0,44	7	7,4
CGTGGA	0,81	0,68	6	3,3
GTGGAA	0,12	1,68	8	0,5
TTCCAC	0,36	1,55	8	-
TCCACG	0,93	0,64	6	-
CCACGG	2,77	0,48	7	-
CACGGA	0,12	0,59	5	-
TCCGCG	0,38	0,52	5	4,5
CCGCGG	4,03	0,48	8	2,6
CGCGGA	0,65	0,45	5	4,5

Pomocí analýzy oligonukleotidů o délce 7 (Tabulka 4.14) byly oba motivy nalezeny s nižšími koeficienty významnosti než v předchozích případech. Oligonukleotidy korespondující s prvním motivem i jeho komplementem dosáhly maximální hodnoty vyšší než 2. Palindromický motiv byl nalezen v rámci dvou oligonukleotidů o koeficientech

významnosti nižších než 1. Nízké koeficienty v porovnání s autory uváděnými byly pravděpodobně způsobeny nižším výskytem jednotlivých oligonukleotidů ve zkoumaných sekvencích.

Tabulka 4.14: Výsledky vyhledávání motivů délky 7 v rodině PDR

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
TCCGTGG	0,73	0,16	4	6,9
CCGTGGA	2,29	0,15	5	
CGTGGAA	0,86	0,30	5	
TTCCACG	1,13	0,27	5	-
TCCACGG	2,26	0,16	5	
CCACGGA	1,24	0,12	4	
TCCGCGG	0,44	0,19	4	5,6
CCGCGGA	0,60	0,17	4	

Na rozdíl od autorů bylo dosaženo nejvyššího koeficientu významnosti pro první motiv v rámci analýzy oligonukleotidů délky 8 (Tabulka 4.15). Konkrétně byla nejvyšší hodnota dosažena pro oligonukleotid CCGTGGAA. I přes to, že jde o nejvyšší hodnotu dosaženou funkcí *OligoAnalysis*, nebyla překročena hodnota uváděná autory metody. Pravděpodobné příčiny této odlišnosti byly uvedeny v rámci popisu výsledků analýzy motivů o délce 6. Obdobných hodnot dosáhly i oligonukleotidy k motivu komplementární. Palindromickému motivu byl vypočten koeficient významnosti 1,17, který je srovnatelný s autory uváděnou hodnotou.

Tabulka 4.15: Výsledky vyhledávání motivů délky 8 v rodině PDR

Motiv	**Výsledky funkce *OligoAnalysis***			**Koeficient významnosti uváděný autory [14]**
	Koeficient významnosti	**Očekávaný výskyt**	**Reálný výskyt**	
TCCGTGGA	1,55	0,07	4	4,2
CCGTGGAA	3,35	0,07	5	
TTCCACGG	3,13	0,08	5	-
TCCACGGA	1,75	0,06	4	
TCCGCGGA	1,17	0,09	4	1,8

Analýzou obou rodin pomocí funkce *OligoAnalysis* bylo dosaženo uspokojivých výsledků. V každé části analýzy byly nalezeny motivy shodující se s motivy, které uvádí autoři metody Oligo-Analysis ve své analýze. Koeficienty významnosti získané analýzou rodiny MET byly v mnoha případech dokonce vyšší, než uvádí autoři. Lepších výsledků bylo dosaženo pro všechny délky oligonukleotidů kromě 5. Vyšší koeficient významnosti určuje větší výjimečnost daného oligonukleotidu. Výsledky analýzy rodiny PDR dosahovaly ve většině případů nižších hodnot, než hodnot uvedených ve studii autorů. Lepší hodnota byla přidělena pouze palindromickému motivu v rámci analýzy oligonukleotidů délky 6. Nižší hodnoty však nepoukazují na chybu algoritmu, ale na možnost odlišnosti kalibračních sekvencí a tedy kalibrací získaných dat, případně na nepatrné odlišnosti ve zkoumaných sekvencích. V rámci algoritmu je důležité, že pro všechny motivy byly získány kladné koeficienty významnosti. Odchylky v kladných hodnotách nemění nic na tom, že může jít o motiv a je potřeba daný oligonukleotid blíže prozkoumat.

ZÁVĚR

V této bakalářské práci byl popsán proces transkripce DNA sekvencí a význam transkripčních motivů. Byly prostudovány tři databáze transkripčních motivů dostupné na internetu a sepsány jejich klady a zápory. Nejobsáhlejší databází je TRANSFAC 7.0, z hlediska přehlednosti pro uživatele jsou však vhodnější databáze MotifMap nebo JASPAR.

V úvodu třetí kapitoly byla uvedena kritéria dělení algoritmů pro vyhledávání transkripčních motivů. Používají se především metody založené na vyhledávání slov a pravděpodobnostní metody. Nejnovější algoritmy jsou kombinací těchto dvou metod. Bývají tedy účinnější, ovšem jejich algoritmy jsou již velmi složité. Následující podkapitoly byly věnovány podrobnému popisu tří algoritmů založených na vyhledávání slov (Oligo-Analysis, Yeast Motif Finder a MDScan) a čtyř algoritmů založených na pravděpodobnostních metodách (MEME, AlignACE, BioProspector a MotifSampler). Každý z uvedených algoritmů má určité klady a zápory. Pro větší soubory vstupních sekvencí je vhodný BioProspector. V případě, že neznáme či nedokážeme odhadnout šířku vyhledávaného motivu je vhodnější algoritmus MEME.

Praktická část této práce je zaměřena na realizaci algoritmu Oligo-Analysis v programovém prostředí Matlab. V prvním úseku je detailně popsáno všech sedm funkcí vytvořených pro tento algoritmus. Mezi hlavní funkce patří *kalibrace* a *OligoAnalysis*, které využívají vedlejší funkce *ctifasta, komplement, kmerpocet* a *kombinace.* Doplňkovou funkcí je *ExpOcc*, jejíž výstup má pro uživatele pouze informativní charakter ohledně očekávaného výskytu jednotlivých oligonukleotidů ve zkoumaných sekvencích.

V rámci dalšího úseku byl popsán proces kalibrace a výběr vhodných dat. Kalibrace je nesmírně důležitou částí celého procesu, neboť na ní závisí výsledky analýzy. Nevhodně zvolené kalibrační sekvence mohou vést ke špatným závěrům. Pro veškeré analýzy v rámci této práce byla použita kalibrační data získaná z nekódujících oblastí genomu kvasinky *Saccharomyces cerevisiae*. Sekvence těchto oblastí byly získány z databáze Saccharomyces Genome Database a byly vstupem pro funkci *kalibrace.*

Ve třetím úseku praktické části je uveden proces validace algoritmu na umělých sekvencích s vnesenými motivy. Sekvence pro validaci byly vytvořeny náhodným rozložením nukleoidů a do těchto sekvencí byly vneseny dva různé motivy, kdy jeden z nich byl palindromem. Motivy byly vyhledávány pro délky 4 až 8 a vnesené motivy byly ve všech případech nalezeny a označeny vysokým koeficientem významnosti. Pro délky oligonukleotidů 6 a větší se koeficient významnosti pohyboval okolo hodnoty 10, což znamená, že bychom očekávali nález jednoho daného oligonukleotidu v rámci 10^{10} rodin koregulovaných genů. Tento fakt dokládá výjimečnost daného oligonukleotidu.

Poslední úsek praktické části práce se věnuje analýze rodin MET a PDR u kvasinky *Saccharomyces cerevisiae*. Tato konkrétní analýza byla vybrána z toho důvodu, že bylo možné její výsledky porovnat s výsledky autorů metody Oligo-Analysis. Analýza rodiny MET pomocí funkce *OligoAnalysis* dosahovala v mnohých případech dokonce lepších výsledků, než uvádí autoři metody, konkrétně při délkách vyhledávaných oligonukleotidů 4, 6, 7 i 8. Motivy nalezené pomocí vytvořené funkce v těchto případech dosahovaly vyššího koeficientu významnosti, tedy větší předpokládané ojedinělosti tolikanásobného výskytu daného oligonukleotidu. Výsledky analýzy rodiny PDR byly v drtivé většině horší, než výsledky autorů metody. Ve všech případech však byly motivy nalezeny a byl jim přidělen kladný koeficient významnosti, což dokládá vyšší výskyt daných oligonukleotidů, než bylo předpokládáno.

Výsledky vytvořených funkcí jsou celkově hodně závislé na počáteční kalibraci. Je tedy možné, že s jinými kalibračními daty by bylo dosaženo lepších výsledků, než uvádí tato práce. Provedené analýzy však dokazují, že algoritmus funguje bez problémů, a dokáže nalézt motivy v sekvencích po správné kalibraci.

LITERATURA

[1] VACÍK, Jiří. *Přehled středoškolské chemie*. 2. vyd. Praha: SPN, 1999, 365 s. ISBN 80-723-5108-7.

[2] IPSER, Jan. *Základy genetiky.* Vyd. 1. Ústí nad Labem: Univerzita J. E. Purkyně, Přírodovědecká fakulta, 2005, 196 s. ISBN 80-704-4707-9.

[3] CLANCY, Suzane. DNA Transcription. *Nature Education* [online]. 2008 [cit. 2014-11-06]. Dostupné z: http://www.nature.com/scitable/topicpage/dna-transcription-426

[4] CAMPBELL, Neil A. a Jane B. REECE. *Biologie.* Vyd. 1. Brno: Computer Press, 2006, xxxiv, 1332 s. ISBN 80-251-1178-4.

[5] KOCHOVÁ, Pavlína. *Metabolismus bílkovin.* [online]. 2013 [cit. 2014-11-06]. Obrázek dostupný z: http://www.mojechemie.cz/images/thumb/Transkripce.png/500px-Transkripce.png

[6] BEDNÁŘ, Jan, Jiří KUCIEL a Tomáš VYHNÁLEK. *Genetika.* Vyd. 2., nezměn. V Brně: Mendelova univerzita, 2010, 148 s. ISBN 978-80-7375-448-8.

[7] LEDVINA, Miroslav, Alena STOKLASOVÁ a Jaroslav CERMAN. *Biochemie pro studující medicíny.* Vyd. 1. Praha: Karolinum, 2004. ISBN 80-246-0851-0.

[8] DAS, M. K. a DAI, H. K. A survey of DNA motif finding algorithms. *BMC Bioinformatics.* 2007, vol. 8, Suppl 7, s. 198-211.

[9] HELDEN, J. v., Alma. F. RIOS a Julio COLLADO-VIDES. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research* [online]. vol. 28, issue 8, s. 1808-1818 [cit. 2014-11-22]. DOI: 10.1093/nar/28.8.1808. Dostupné z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/28.8.1808

[10] MATHELIER, A., ZHAO, X., ZHANG, A. W., PARCY, F., WORSLEY-HUNT, R., ARENILLAS, D. J., BUCHMAN, S., CHEN, C.-y., CHOU, A., IENASESCU, H., LIM, J., SHYR, C., TAN, G., ZHOU, M., LENHARD, B., SANDELIN, A. a WASSERMAN, W. W. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* [online]. 2013 [cit. 2014-12-04]. Dostupné z: http://jaspar.genereg.net/

[11] MATHELIER, A., ZHAO, X., ZHANG, A. W., PARCY, F., WORSLEY-HUNT, R., ARENILLAS, D. J., BUCHMAN, S., CHEN, C.-y., CHOU, A., IENASESCU, H., LIM, J., SHYR, C., TAN, G., ZHOU, M., LENHARD, B., SANDELIN, A. a

WASSERMAN, W. W. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. [online]. *Nucleic Acids Research* [online]. 2013 [cit. 2014-12-10]. Obrázek dostupný z: http://jaspar.genereg.net/cgi-bin/jaspar_db.pl?ID=MA0133.1&rm=present&collection=CORE

[12] MATYS, V., KEL-MARGOULIS, O. V., FRICKE, E., LIEBICH, I., LAND, S., BARRE-DIRRIE, A., REUTER, I., CHEKMENEV, D., KRULL, M., HORNISCHER, K., VOSS, N., STEGMAIER, P., LEWICKI-POTAPOV, B., SAXEL, H., KEL, A. E. a WINGENDER, E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* [online]. 2006 [cit. 2014-12-04]. Dostupné z: http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/search.cgi

[13] DAILY, Kenneth, PATEL, Vishal R., RIGOR, Paul, XIE, Xiaohui a Pierre BALDI. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics* [online]. 2011 [cit. 2014-12-04]. Dostupné z: http://motifmap.ics.uci.edu/

[14] HELDEN, J. van, B. ANDRÉ a J. COLLADO-VIDES. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*. 1998, vol. 281, issue 5, s. 827-842. DOI: 10.1006/jmbi.1998.1947.

[15] SINHA, Saurabh a Martin TOMPA. A statistical method for finding transcription factor binding site. *Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology*. 2000.

[16] HU, J., B. LI a D. KIHARA. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research* [online]. 2005-09-07, vol. 33, issue 15, s. 4899-4913 [cit. 2014-12-29]. DOI: 10.1093/nar/gki791. Dostupné z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gki791

[17] LIU, X. Shirley, Douglas L. BRUTLAG a Jun S. LIU. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* [online]. 2002-7-8, vol. 20, issue 8, s. 835-839 [cit. 2014-12-29]. DOI: 10.1038/nbt717. Dostupné z: http://www.nature.com/doifinder/10.1038/nbt717

[18] BAILEY, Timothy L. a Charles ELKAN. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* [online]. 1995, vol.

21, 1-2, s. 51-80 [cit. 2014-12-30]. DOI: 10.1007/BF00993379. Dostupné z: http://link.springer.com/10.1007/BF00993379

[19] LAWRENCE, C., S. ALTSCHUL, M. BOGUSKI, J. LIU, A. NEUWALD a J. WOOTTON. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* [online]. 1993-10-08, vol. 262, issue 5131, s. 208-214 [cit. 2014-12-31]. DOI: 10.1126/science.8211139. Dostupné z: http://www.sciencemag.org/cgi/doi/10.1126/science.8211139

[20] ROTH, Frederick P., Jason D. HUGHES, Preston W. ESTEP a George M. CHURCH. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* [online]. 1998, vol. 16, issue 10, s. 939-945 [cit. 2014-12-31]. DOI: 10.1038/nbt1098-939. Dostupné z: http://www.nature.com/doifinder/10.1038/nbt1098-939

[21] LIU, X., D. L. BRUTLAG, J. S. LIU a George M. CHURCH. BIOPROSPECTOR: DISCOVERING CONSERVED DNA MOTIFS IN UPSTREAM REGULATORY REGIONS OF CO-EXPRESSED GENES. *Biocomputing 2001* [online]. WORLD SCIENTIFIC, 2000, vol. 16, issue 10, s. 127-138 [cit. 2014-12-31]. DOI: 10.1142/9789814447362_0014. Dostupné z: http://www.worldscientific.com/doi/abs/10.1142/9789814447362_0014

[22] THIJS, Gert, Kathleen MARCHAL, Magali LESCOT, Stephane ROMBAUTS, Bart DE MOOR, Pierre ROUZÉ a Yves MOREAU. A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes. *Journal of Computational Biology* [online]. 2002, vol. 9, issue 2, s. 447-464 [cit. 2015-01-01]. DOI: 10.1089/10665270252935566. Dostupné z: http://www.liebertonline.com/doi/abs/10.1089/10665270252935566

SEZNAM ZKRATEK

DNA	Deoxyribonukleová kyselina
RNA	Ribonukleová kyselina
mRNA	Mediátorová ribonukleová kyselina
A	Adenin
C	Cytosin
T	Thymin
G	Guanin
bp	base pairs (párů bází)
ORF	Open reading frame (otevřený čtecí rámec)
YMF	Yeast Motif Finder

PŘÍLOHA A: PSEUDOKÓD ALGORITMU

OligoAnalýza(kalibrační sekvence, délka motivu, sekvence)

1 oligonukleotidy ← vypsat kombinace (délka motivu)

2 předpokládané frekvence ← spočítat frekvence (kalibrační sekvence, délka motivu)

3 počet sekvencí ← spočítat řádky (kalibrační sekvence)

4 počet oligonukleotidů ← spočítat sloupce (oligonukleotidy)

5 for a = 1:počet sekvencí

délka sekvence (a) ← spočítat sloupce (kalibrační sekvence (a))

shodné pozice = délka sekvence (a) – délka motivu + 1

shodné pozice celkem = shodné pozice celkem + shodné pozice

6 shodné pozice celkem = 2 * shodné pozice celkem

7 if délka motivu/2 je celé číslo

počet palindromů = 4^(délka motivu/2)

8 else

počet palindromů = 0

9 odlišné oligonukleotidy = 4^(délka motivu) – (4^(délka motivu) – počet palindromů)/2

10 for b = 1:počet oligonukleotidů

počet výskytů (oligonukleotidy (b)) ← spočítat výskyt (sekvence)

pravděpodobnost výskytu (oligonukleotidy (b)) ← vypočítat pravděpodobnost (shodné pozice celkem, počet výskytů (oligonukleotidy (b)), předpokládané frekvence (b))

koeficient významnosti ← -log(pravděpodobnost výskytu (oligonukleotidy (b)) * odlišné oligonukleotidy)

PŘÍLOHA B: VÝVOJOVÝ DIAGRAM ALGORITMU